DTS311TC FINAL YEAR PROJECT

FusionMedCLIP: A Framework for Few-Shot Medical Anomaly Detection via Adapted Vision-Language Model

In Partial Fulfillment
of the Requirements for the Degree of
Bachelor of Engineering

By

Ziqian Zhang
ID 2144337

Supervisors Name

Dr. Kang Dang
Dr. Bin Chen

School of AI and Advanced Computing
XI'AN JIAOTONG-LIVERPOOL UNIVERSITY

April 2025

Abstract

The integration of artificial intelligence into medical imaging holds great promise for improving diagnostic accuracy and efficiency. Among these applications, detecting anomalies in medical images is crucial for timely diagnosis, yet accurately identifying abnormalities with limited labelled data presents a significant challenge due to the inherent rarity of anomalies and the high cost of expert annotation. This necessitates effective few-shot learning approaches. While large Vision-Language Models, such as CLIP, possess strong generalisation capabilities, their direct application to specialised medical domains is hampered by domain shift and the difficulty of crafting effective textual prompts.

To address these issues, this work introduces FusionMedCLIP, a framework designed to efficiently adapt large vision-language models for few-shot medical anomaly detection through synergistic adaptation. The key **4** contributions lie in the synergistic co-optimization of several components: **(i)** initial medical domain alignment via ROCO pre-training, **(ii)** automated, knowledge-augmented prompt generation (CoOp+UMLS) replacing manual crafting, **(iii)** lightweight, parameter-efficient fine-tuning of the visual encoder (ViT) using LoRA and Block Adapters, and **(iv)** pathology-guided anomaly synthesis to effectively utilize limited training data. This integrated strategy specifically targets the challenges of few-shot learning in clinical settings. Validated on benchmarks including BrainMRI, BUSI, and CheXpert, FusionMedCLIP achieves state-of-the-art few-shot performance, reaching up to 94.8% Image-AUROC on BrainMRI and 89.9% Pixel-AUROC on BUSI (k=16). This significantly surpasses zero-shot CLIP (>37% Image-AUROC gain on BrainMRI) and leading external anomaly detection methods. By synergistically adapting, FusionMedCLIP offers a practical and effective pathway to leverage their power for medical anomaly detection, specifically addressing the dual challenges of domain adaptation and data scarcity in low-resource clinical settings.

Keywords: Medical Anomaly Detection, Few-Shot, CLIP, Vision-Language Models, Parameter-Efficient Tuning, Prompt Learning, Domain Adaptation

Acknowledgements

I would like to express my gratitude to Dr. Kang Dang for his insightful guidance at the project's outset. His profound understanding of medical anomaly detection offered essential background, shaping the scope and direction of this research and contributing significantly to the formulation of a focused research proposal.

I am also sincerely indebted to Dr. Bin Chen for his sustained academic mentorship. His critical feedback on the technical methodology, experimental design, and model refinement proved invaluable in navigating research challenges. His advice on manuscript's structure and presentation was instrumental in improving the clarity and rigour of this dissertation.

This research was undertaken as part of the final year undergraduate studies at Xi’an Jiaotong-Liverpool University (XJTLU). I acknowledge the supportive academic environment and resources provided by XJTLU, which were conducive to this work.

Contents

List of Tables

List of Figures

List of Acronyms

AD	Anomaly Detection
AE	Autoencoder
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
CoOp	Context Optimization
CT	Computed Tomography
ZSL	Zero-Shot Learning
FSL	Few-Shot Learning
GAN	Generative Adversarial Network
KG	Knowledge Graph
LLRD	Layer-wise Learning Rate Decay
LoRA	Low-Rank Adaptation
MAML	Model-Agnostic Meta-Learning
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
PEFT	Parameter-Efficient Fine-Tuning
PET	Positron Emission Tomography
QKV	Query, Key, Value (in Transformer Attention)
ROCO	Radiology Objects in Context
SSL	Self-Supervised Learning
SOTA	State-of-the-Art
UMLS	Unified Medical Language System
VAE	Variational Autoencoder
ViT	Vision Transformer
VLMs	Vision-Language Models
ROCO-CLIP	CLIP model fine-tuned on the ROCO dataset

Chapter 1
Introduction

1.1 Background

Medical image analysis is a cornerstone of modern precision medicine, providing vital tools for early disease detection, accurate diagnosis, and personalised treatment planning [1][2]. Within this critical field, **Anomaly Detection (AD)** represents a foundational task, focusing on the automated identification of patterns in medical scans that deviate significantly from established normalcy in medical scans. Recognising such anomalies, which may include subtle lesions, nascent tumours, or other unexpected pathological changes, is crucial for timely clinical intervention and improved patient outcomes.

However, the development and practical deployment of effective medical AD systems face several significant hurdles. Firstly, a foremost problem faced is the **scarcity of well-characterised anomaly data** for training. While normal medical images may be relatively abundant, specific abnormalities, particularly those associated with rare diseases or subtle variations, are inherently infrequent [8]. This makes assembling large, representative datasets of labelled abnormal examples exceptionally difficult. Secondly, acquiring the necessary supervisory input, even for a small number of abnormal cases, poses difficulties. Whether requiring simple confirmation labels, precise localisation information, or **clinically accurate textual descriptions** (as needed for vision-language models), the process demands considerable time from domain experts (e.g., radiologists) and is consequently resource-intensive [3]. Thirdly, the intrinsic **heterogeneity and ambiguity of anomalies** themselves; they can vary widely in appearance, size, and location, sometimes manifesting as only minor deviations from the norm, making robust generalisation challenging [7]. Lastly, stringent **privacy regulations** governing medical data (e.g., GDPR, HIPAA) restrict large-scale data sharing and centralized aggregation, further limiting the scale of data available for model development [15].

These converging challenges, particularly the paucity of labelled anomaly instances and the cost associated with obtaining expert supervision, render traditional supervised learning methods, which rely on extensive labelled datasets, often impractical for medical AD. This reality strongly motivates the exploration of alternative learning paradigms capable of operating effectively under severe data constraints. Consequently, **Few-Shot Learning (FSL)**, which focuses on learning to recognise new categories (in this case, anomalies) from only a handful of labelled examples or limited guidance, has emerged as an essential research frontier, holding significant promise for advancing reliable and scalable diagnostic support in real-world clinical settings [16].

1.2 Limitations of Existing Methods (Motivation)

While the need for effective few-shot medical anomaly detection is clear, current approaches face major limitations, driving the need for new strategies. Traditional supervised deep learning methods, mainly classification and segmentation [16], are often used in medical image analysis but, as shown in Table 1, they require large, carefully labelled datasets for predefined conditions or specific structures. This requirement often clashes with the practical limits in medical imaging, where obtaining these extensive annotations, especially for varied abnormal findings, is often very difficult and expensive. The Anomaly Detection (AD) approach [2][10], in contrast, takes a different path, focusing primarily on identifying differences from what is learned as normal, instead of recognising many specific, predefined classes. Although this goal might, in theory, reduce the need for complete labelling for all possible conditions, using AD effectively, especially in FSL situations, runs into its own key difficulty. As shown in Table 1, this difficulty is often in obtaining and effectively using enough high-quality supervision—whether this means classification labels for the few abnormal samples, precise boundary annotations, or, importantly for modern methods, useful text descriptions—for the naturally rare, varied, and often subtle abnormal examples. This problem of lacking supervision specifically for abnormalities makes it difficult to use even the AD approach effectively given the usual data limits found in clinical practice, highlighting the need for special FSL solutions.

Feature	*Classification*	*Segmentation*	*Anomaly Detection*
Primary Objective	Assign predefined category labels	Delineate specific, known structures	Identify patterns deviating significantly from normality
Input Data Focus	Labelled examples for each target class	Data with pixel-level annotations	Primarily normal data; focus on leveraging scarce abnormal samples/signals effectively
Supervision Requirement	High (Large labelled datasets)	Very High (Dense pixel annotations)	Varies: Low (unsupervised) to Moderate (semi-supervised, few-shot, VLM-based needing text)
Output Example	Disease type (e.g., 'Cancerous')	Tumour mask, organ boundaries	Anomaly score, binary flag ('normal'/'abnormal'), anomaly map
Key Challenges	Class imbalance, fine-grained diff	Annotation cost, boundary ambiguity	Defining 'normality', heterogeneity of unforeseen anomalies, **supervision scarcity** for abnormal class

Table 1. Comparative Analysis of Characteristics in Classification, Segmentation, and Anomaly Detection Tasks.

Beyond this fundamental challenge of scarce supervision for abnormal cases, conventional deep learning models adapted for AD often exhibit poor generalisation and are prone to overfitting when trained with limited data [2][16]. Recognizing this bottleneck, research has explored various FSL strategies, achieving notable success in general computer vision domains through techniques like metric learning [23], meta-learning [24], and generative modelling [4][8]. Yet, the direct transplantation of these methods to medical imaging often falls short. A primary reason is the substantial domain gap: medical images possess unique characteristics (e.g., grayscale nature, complex textures, subtle pathological indicators) that differ markedly from the natural images used for pre-training general vision models, often requiring finer-grained analysis [6]. Moreover, many established FSL methods operate solely on visual data, failing to exploit the rich multimodal context intrinsic to clinical practice, such as information contained within radiology reports or codified medical knowledge, which could supply valuable semantic priors. Conversely, emerging text-guided diagnostic approaches can struggle with the practical need for consistently precise and standardised pathological descriptions, which are difficult to acquire reliably, especially for rare or newly characterised anomalies [11].

In this context, large-scale Vision-Language Models (VLMs) pre-trained on web-scale data, exemplified by CLIP [1], have emerged as a potent foundation. CLIP's ability to learn robust, joint representations of images and text facilitates remarkable zero-shot and few-shot generalisation across diverse visual tasks, making it highly attractive for tackling few-shot medical AD. However, realizing this potential is non-trivial, and directly applying off-the-shelf CLIP models encounters critical hurdles. Firstly, the significant domain mismatch between CLIP's pre-training data (general web images/text) and the specialized medical domain necessitates careful domain adaptation [7][13]. Secondly, CLIP's effectiveness heavily relies on hand-crafted text prompts; designing optimal prompts for specific, often nuanced, medical anomalies is a challenging and inflexible process known as prompt engineering [11]. Thirdly, the pre-trained visual representations, while powerful, might not be inherently optimized for the fine-grained nature of many medical abnormalities, indicating a need for targeted visual feature alignment [6]. Finally, performing full fine-tuning on these massive models is computationally prohibitive and risks overfitting in few-shot scenarios, contradicting the need for parameter-efficient adaptation strategies.

1.3 Methods and Contributions (Aims & Objective)

Addressing the limitations outlined above, this work introduces **FusionMedCLIP**, a novel framework designed to enhance and integrate the abilities of CLIP [1] for the crucial objective of **few-shot** medical image AD. Recognizing the need for both domain specificity and parameter efficiency, FusionMedCLIP employs a multi-faceted strategy. First, to mitigate the domain discrepancy between natural web data and clinical imagery, I leverage the diverse **ROCO dataset** [17] for comprehensive **pre-finetuning** of CLIP, establishing a robust medical vision-language foundation applicable across various modalities and conditions. Second, confronting the challenges of manual prompt engineering and the need for nuanced medical descriptions, I incorporate knowledge-enhanced learnable prompts through **CoOp** [11] integrated with the **UMLS** [18] medical knowledge base. This enables the model to automatically learn clinically relevant semantic representations for anomalies, guided by standardized medical ontologies, thereby reducing dependence on handcrafted or potentially unavailable expert textual inputs. Third, to effectively adapt the visual representations without catastrophic forgetting or prohibitive computational cost, I propose a synergistic parameter-efficient tuning approach. This combines **LoRA** [19], applied globally to key layers (Attention QKV, MLP) within the **ViT-L/14** [14] encoder, with lightweight **Adapters** specifically targeting the outputs of intermediate layers crucial for interaction with the learned prompts. This strategy facilitates precise visual feature refinement and effective vision-language alignment in an end-to-end manner, while maintaining computational tractability. Finally, acknowledging the pervasive issue of scarcity of annotated anomalies in clinical practice, FusionMedCLIP integrates **multi-method anomaly synthesis** during training, generating varied pseudo-anomalous examples to bolster model robustness and generalization from limited real abnormal data. Establishing state-of-the-art performance on multiple challenging few-shot medical anomaly detection benchmarks (e.g., Brain MRI [20], BUSI [21], CheXpert [22]).

The core contributions of **FusionMedCLIP** are thus multifaceted:

1. Proposing **FusionMedCLIP**, an integrated framework synergizing domain pre-finetuning, knowledge-guided prompt learning, hybrid parameter-efficient tuning, and anomaly synthesis for effective few-shot medical AD.
2. Demonstrating the critical necessity of broad **medical domain pre-finetuning** (on ROCO) for adapting large VLMs to specialized medical imaging tasks.
3. Introducing a **CoOp+UMLS** mechanism that automates medically relevant prompt generation, significantly reducing reliance on manual engineering.
4. Using a **synergistic parameter-efficient tuning** strategy (LoRA+Adapters) enabling computationally efficient yet precise visual feature adaptation for medical AD.

The remainder of this dissertation is organized as follows: Section 2 reviews related work. Section 3 details proposed FusionMedCLIP. Section 4 presents the experimental setup and results. Finally, Section 5 concludes the dissertation and discusses future directions.

Chapter 2

Literature Review

2.1 Conventional Medical Anomaly Detection

Supervised deep learning models, primarily Convolutional Neural Networks (CNNs) and segmentation architectures like the U-Net [25][26], have historically been effective for specific medical anomaly detection tasks when large labeled datasets are available [5]. Their strength lies in learning direct mappings to semantic labels for known anomaly types. However, their practical utility is often challenged by the significant cost and expertise needed for creating comprehensive annotations, a major bottleneck in the medical field [16]. This dependency makes them prone to overfitting and limits their generalization capabilities, especially for rare conditions or when faced with limited training samples [24].

To mitigate the reliance on extensive annotations, unsupervised and self-supervised learning (SSL) methods have emerged as prominent alternatives. Unsupervised techniques typically learn a model of 'normality' from unlabeled healthy data, identifying anomalies as deviations, for instance, through reconstruction errors using Autoencoders or Generative Adversarial Networks (GANs) [4]. Self-supervised learning seeks to learn rich feature representations from unlabeled data via pretext tasks [6]. However, despite reducing the annotation burden, both conventional unsupervised and SSL approaches face inherent limitations in the context of medical anomaly detection. They often exhibit reduced sensitivity to subtle anomalies that closely resemble normal patterns and can be confounded by the high degree of natural anatomical and imaging variability inherent in medical data [12]. Furthermore, these methods typically yield anomaly scores or localization maps without providing semantic information about the nature of the abnormality (i.e., what kind of anomaly it is). Critically, their inability to effectively capitalize on limited available supervision, which is often the practical scenario in clinics, remains a significant drawback [16].

Paradigm	*Core Principle*	*Advantages*	*Key Limitations*	*Methods*
Supervised Learning	Learn mapping from images to labels using annotations	High accuracy for known anomalies with sufficient data; Direct semantic out	Heavy reliance on large, expensive annotated datasets; Poor generalization under data scarcity; Struggles with novel anomalies	CNNs, U-Net
Unsupervised / SSL	Learn patterns from unlabeled data (normality / features)	Reduces annotation burden; Potential to detect novel anomalies	Sensitivity to subtle anomalies; High false positive rates; Lack of semantic understanding; Difficulty leveraging limited labels	Reconstruction (AE, GANs)

Table 2. Comparison of Conventional Paradigms for Medical Anomaly Detection

The distinct advantages and disadvantages characterizing these traditional supervised and unsupervised/SSL paradigms, **as comparatively outlined in Table 2**. Challenges such as data scarcity, the need for semantic understanding, and the difficulty of leveraging sparse annotations highlight a critical need for approaches specifically designed to operate effectively under severe data constraints.

2.2 Strategies for Limited Medical Data

To directly confront the data scarcity prevalent in medical anomaly detection, Few-Shot Learning (FSL) and Zero-Shot Learning (ZSL) paradigms have garnered significant interest [16]. FSL aims to enable models to learn effectively from a very small number of labeled examples per class, while ZSL targets the recognition of classes not encountered during the training phase, often leveraging auxiliary semantic information. Various approaches within these paradigms have been explored for medical applications, each with distinct mechanisms and associated challenges, as summarized in Table 3.

Several FSL paradigms actively tackle data scarcity in medicine. **Metric learning** methods, including Prototypical Networks [27], learn embedding spaces where classification relies on distance to class prototypes derived from the few available examples [23]. While conceptually straightforward, defining robust prototypes is inherently difficult due to

Paradigm	*Core Principle*	*Advantages*	*Key Limitations*	*Methods*
Metric Learning	Learn embedding space; classify by distance to prototypes/support set	Conceptually simple; Direct comparison	Sensitive to high intra-class variance & subtle differences; Defining robust prototypes from few samples is hard	Prototypical Nets
Meta-Learning	Learn adaptable model initializations ("learning to learn")	Fast adaptation potential	Optimization complexity/instability; May underperform on complex, fine-grained adaptation	MAML
Generative / Aug.	Synthesize additional training data from few examples	Data augmentation for scarce classes	Realism and diversity of synthesized medical data; Potential for artifacts	GAN-based synthesis

Table 3. Overview of FSL Paradigms in Medical Imaging

the significant intra-class variation (e.g., diverse pathology appearances) and subtle inter-class differences commonly encountered in medical images, which can limit discriminative power [13]. **Meta-learning** strategies, exemplified by MAML [28], focus on acquiring highly adaptable model initializations by training across various simulated few-shot tasks ("learning to learn"). Although promoting rapid adaptation, they often entail complex optimization routines, potential instability, and the learned prior might lack the specificity needed for precise adaptation to nuanced, fine-grained medical anomaly detection tasks using extremely sparse data [16][24]. Furthermore, **generative approaches** [4] attempt to augment the limited datasets by synthesizing supplementary training instances, frequently employing GANs. The effectiveness of this strategy critically hinges on the ability to generate synthetic data possessing both high fidelity (capturing essential textures and structural details) and sufficient diversity (representing the relevant spectrum of variations). Achieving this for complex medical images remains a significant challenge, with risks of unrealistic outputs potentially impairing rather than aiding model training [8]

While existing FSL and ZSL paradigms offer strategies to address data scarcity, their direct application to medical image anomaly detection often falls short of clinical requirements. This is attributed to inherent limitations such as domain specificity, the subtle nature of anomalies, insufficient utilization of semantic information, and difficulties in integrating prior knowledge. These challenges necessitate the development of novel methods better adapted to the unique characteristics of medical imaging, capable of effectively leveraging multimodal information (e.g., text), and efficiently utilizing limited annotated data.

2.3 Vision-Language Models in Medicine

The advent of large-scale Vision-Language Models (VLMs), spearheaded by CLIP [1], represents a significant shift in leveraging multimodal data. Pre-trained on vast datasets of internet-sourced image-text pairs using a contrastive objective, CLIP learns joint embeddings that align visual concepts with their natural language descriptions. This alignment endows CLIP with remarkable zero-shot and few-shot generalization capabilities on diverse downstream tasks, particularly in natural image domains, simply by evaluating image-text similarity against prompted class descriptions.

Mathematically, CLIP employs two separate encoders: an image encoder $f_\theta(I)$ and a text encoder $g_\phi(T)$, where I represents an input image and T denotes its associated textual description [14][3][5]. The parameters θ and ϕ are learned during training. Both encoders project their inputs into a shared latent space of dimension d, producing normalized embedding vectors $\mathbf{v}_I$ and $\mathbf{v}_T$.

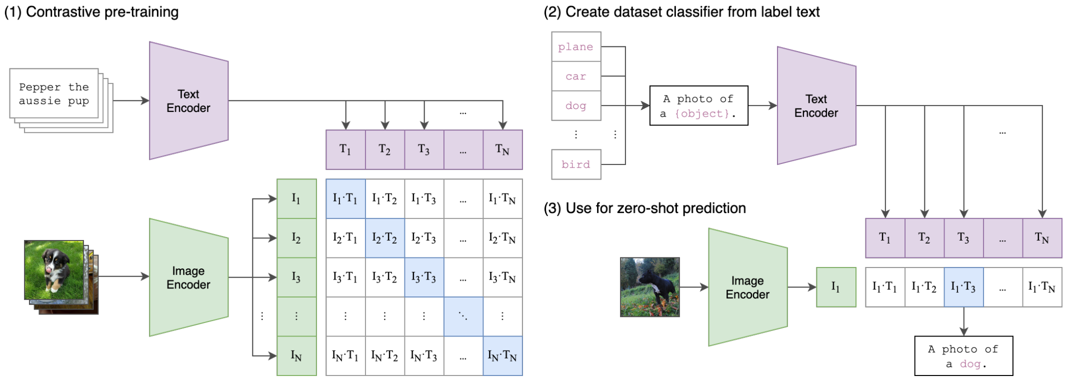
(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ⋮ IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
T1 T2 T3 ... TN
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Figure 1. An overview of the CLIP model's image-text alignment mechanism [1]. On the left, the model learns joint embeddings by computing the similarity between image features and textual prompts across all categories. On the right, the learned representations are used to predict the most semantically similar textual label for a given image, enabling zero-shot classification.

The training objective is to maximize the cosine similarity between embeddings of matched image-text pairs while minimizing it for mismatched pairs [11]. This is formalized using a contrastive loss function. $\mathrm{sim}(\mathbf{v}_I, \mathbf{v}_T) = \mathbf{v}_I^{\top}\mathbf{v}_T$ represents the cosine similarity between the image and text embeddings, and τ is a temperature hyperparameter that controls the concentration level of the distribution. In the anomaly detection phase, CLIP-based methods adapt this framework by utilizing textual descriptions of normal and abnormal conditions [13]. For a medical image I, the image encoder generates its embedding $\mathbf{v}_I$.

This inherent potential for zero- and few-shot learning naturally spurred explorations into adapting CLIP for medical imaging tasks, aiming to mitigate the data scarcity issues discussed earlier. Initial studies demonstrated promising, albeit sometimes limited, zero-shot classification performance on tasks such as chest X-ray interpretation [5][7] and pathology image analysis using straightforward text prompts. Concurrently, research probed the transferability and utility of off-the-shelf CLIP visual features for various medical applications, often finding them beneficial but not universally optimal compared to domain-specific models [6][16].

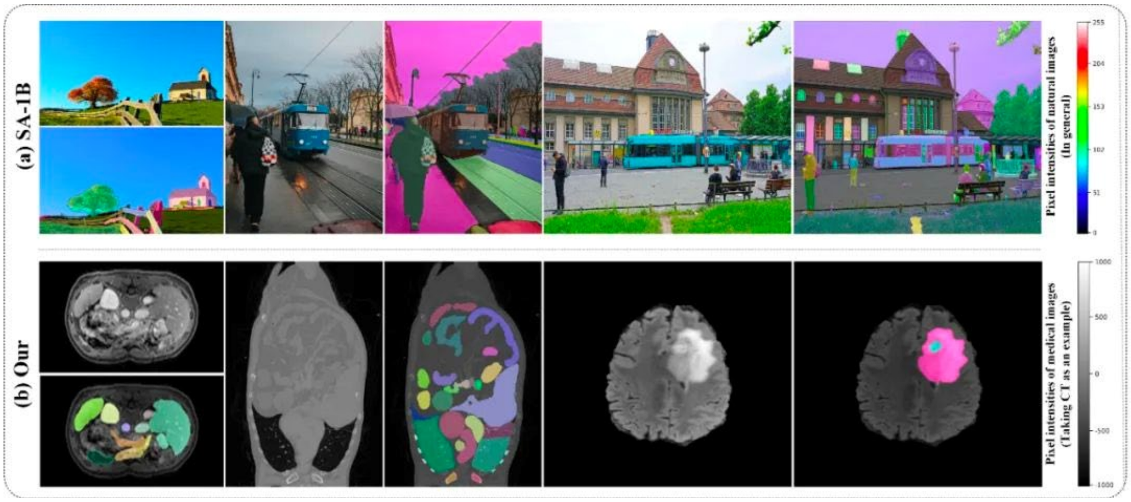
(a) SA-1B
(b) Our
Pixel intensities of natural images (In general)
255
204
153
102
51
0
Pixel intensities of medical images (Taking CT as an example)
1000
500
0
-500
-1000

Figure 2. Characteristic Difference Map: Nature image (top) vs. medical image (bottom). The figure shows the differences in color, structural complexity and texture features between natural and medical images, and the impact of these differences on CLIP encoders, highlighting the problem of domain migration.

However, directly applying CLIP models pre-trained on general web data to the medical domain encounters a significant performance drop, often referred to as the domain gap [3]. Several factors contribute to this disparity. First, medical images (e.g., grayscale MRI/CT, radiographs, histopathology slides) possess distinct visual characteristics, including subtle textures, complex anatomical structures, and lack of color cues prevalent in natural images, deviating significantly from CLIP's pre-training distribution [17]. Second, medical language exhibits high specificity, employing specialized terminology and contextual nuances absent in the general web text used for CLIP's pre-training [6]. Third, medical tasks frequently demand fine-grained analysis, such as identifying small lesions or subtle textural changes, requiring a level of detail sensitivity potentially underdeveloped by CLIP's training on broader object recognition. Table 3 provides a summary of these strengths and challenges.

	Aspect	*Description*	*Examples*
Strengths	Knowledge Transfer	Leverages pretraining on diverse datasets to generalize well in few-shot scenarios.	Fine-tuning with a small dataset of chest X-rays enables generalization to pneumonia detection.
	Prompt Flexibility	Allows integration of expert knowledge through textual prompts.	"A chest X-ray showing signs of pneumonia" versus "a chest X-ray of a healthy lung."
Challenges	Domain Shift	Medical images differ from natural images in grayscale intensity, anatomical structures, and artifacts.	Chest X-rays lack the color and context present in natural images used for CLIP’s pretraining.
	Medical Semantics	Clinical language is highly specialized and complex, requiring adaptation of the text encoder.	Terms like "ground-glass opacity" or "hyperintensity" are crucial but domain-specific.
	Explain ability	Lacks mechanisms for identifying specific regions contributing to anomaly detection results.	Unable to highlight a tumor location in an MRI despite correctly identifying it as abnormal.

Table 4. The Strengths and Challenges of CLIP-Based Methods in Medical Imaging

2.4 Key Optimization for Medical CLIP Adaptation

Medical Adaptation of VLMs: Pre-training/fine-tuning CLIP on medical datasets (e.g., ROCO) improves domain performance over zero-shot [29]. However, this requires significant resources, and full fine-tuning can dilute generalization while still needing task-specific adjustments. *Gap: Need to parameter-efficient methods for diverse few-shot anomaly detection tasks, rather than relying on full fine-tuning.*

Prompt Optimization: CLIP's reliance on hand-crafted text prompts poses a significant bottleneck, especially for complex medical semantics. While learnable prompts like CoOp [11] automate this process, they typically lack mechanisms to explicitly integrate crucial structured medical knowledge, limiting their semantic grounding in this specialized domain [6][18]. *Gap: Need for knowledge learnable prompts tailored for medicine.*

Parameter-Efficient Fine-Tuning (PEFT): Fully fine-tuning large VLMs like CLIP is computationally expensive and risks catastrophic forgetting in few-shot settings [19]. PEFT methods such as Adapters and LoRA offer solutions by tuning only a small parameter subset [29]. However, existing applications often lack targeted strategies combining PEFT methods synergistically or applying them specifically to layers critical for visual-prompt alignment in medical tasks. *Gap: Need for targeted, synergistic PEFT strategies optimized for few-shot medical VLM adaptation.*

Data Augmentation via Synthesis: Few-shot learning inherently suffers from limited data diversity. While anomaly synthesis techniques (using GANs [4], Diffusion Models, or simpler methods [9]) can augment data, their application to specifically enhance VLM-based few-shot medical anomaly detection is nascent. Critically, leveraging synthesis in a multi-method manner to improve generalization rather than fitting to specific artifacts requires dedicated investigation [8]. *Gap: Need for systematic integration of multi-method anomaly synthesis within few-shot VLM frameworks for medicine.*

Chapter 3

Methodology

This work introduces FusionMedCLIP, a novel framework designed specifically for few-shot AD in medical images via systematically improving and adapting the pre-trained CLIP model. Recognizing the limitations of applying general VLMs directly to specialized medical tasks under data scarcity, FusionMedCLIP adopts a multi-stage optimization strategy. This strategy aims to achieve robust performance through synergistic improvements in domain adaptation, prompt formulation, visual feature refinement, and data resilience, while prioritizing parameter efficiency.

The overall architecture of FusionMedCLIP, depicted in Figure 3, is composed of four interconnected stages executed sequentially or concurrently during training:

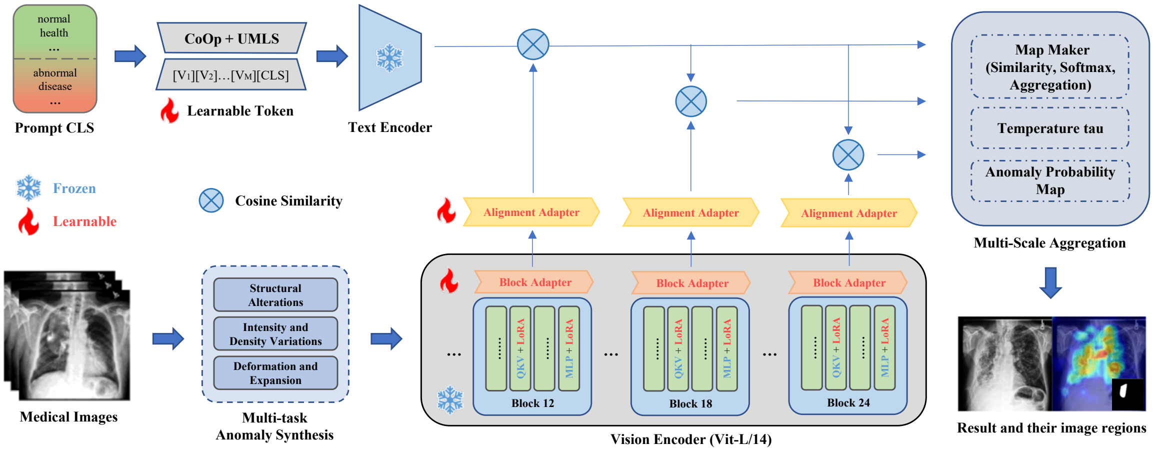

Figure 3. FusionMedCLIP Workflow for FSL Medical Anomaly Detection. Medical images, optionally augmented via Multi-task Anomaly Synthesis, are processed by the Vision Encoder ViT-L/14 (Frozen core, Learnable LoRA/Block Adapters). These features are then fed to Learnable Alignment Adapters for channel alignment and normalization. CoOp+UMLS module generates Learnable prompts processed by the Frozen Text Encoder. Pixel-wise Cosine Similarity is computed between aligned visual features and text embeddings. The Map Maker aggregates multi-scale similarity maps (using a Learnable Temperature) into a final Anomaly Probability Map, driving the optimization of all learnable components via a combined loss.

3.1 Pretraining CLIP

The first stage of FusionMedCLIP focuses on establishing a robust, medically-aware visual foundation. This involves adapting the CLIP ViT-L/14 visual encoder [1][14] using the large-scale, diverse ROCO dataset [17], which contains a wide array of medical images paired with textual captions. The objective is to imbue the visual encoder with general medical domain knowledge before task-specific few-shot fine-tuning.

Based on extensive preliminary experiments evaluating various adaptation strategies (including parameter-efficient methods like LoRA, partial freezing, text encoder modifications [30], and cascaded fine-tuning), I determined that full parameter fine-tuning of only the ViT-L/14 at this stage yielded the effective foundational representations for subsequent downstream adaptation. I hypothesize this approach allows the visual backbone to comprehensively learn relevant medical visual patterns present in ROCO, avoiding potential information flow bottlenecks associated with partial freezing, while targeted parameter-efficient tuning is reserved for the subsequent few-shot stages.

The fine-tuning process optimizes the alignment between image and text representations using the standard CLIP contrastive learning objective [1]. Let $\mathbf{f}_\theta(I)$ be the output of the image encoder for image I, and $\mathbf{g}_\phi(T)$ be the output of the text encoder for text T. The normalized embeddings are computed as:

$$\mathbf{v}_I = \frac{\mathbf{f}_\theta(I)}{\|\mathbf{f}_\theta(I)\|_2}, \mathbf{v}_T = \frac{\mathbf{g}_\phi(T)}{\|\mathbf{g}_\phi(T)\|_2} \tag{1}$$

where $\|\cdot\|_2$ denotes the L2 norm. The similarity between an image embedding and a text embedding is measured by the cosine similarity: $\mathrm{sim}(\mathbf{v}_I, \mathbf{v}_T) = \mathbf{v}_I^T \mathbf{v}_T$. Given a batch of N image-text pairs $\{(I_i, T_i)\}_{i=1}^N$ from ROCO, the model is optimized using a symmetric contrastive loss. This loss function seeks to increase the similarity between the matched image-text pairs (I_i, T_i) while minimizing the similarity of $N^2 - N$ mismatched pairs within the batch. This is formulated as the sum of two cross-entropy losses across the batch similarities:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{\exp(\text{sim}(\mathbf{v}_{I_i}, \mathbf{v}_{T_i})/\tau)}{\sum_{j=1}^{N}\ \exp(\text{sim}(\mathbf{v}_{I_i}, \mathbf{v}_{T_j})/\tau)} + \log\frac{\exp(\text{sim}(\mathbf{v}_{I_i}, \mathbf{v}_{T_i})/\tau)}{\sum_{j=1}^{N}\ \exp(\text{sim}(\mathbf{v}_{I_j}, \mathbf{v}_{T_i})/\tau)}\right) \quad (2)$$

where τ is a temperature parameter that scales the logits [1]. Note that during this stage, only the parameters θ of the vision encoder $\mathbf{f}_{\theta}$ are updated, while the text encoder parameters ϕ might be kept frozen or adapted separately depending on the specific setup.

To ensure robust training and mitigate overfitting during this full fine-tuning phase using $\mathcal{L}_{\text{CLIP}}$, I employed a combination of advanced optimization and regularization techniques. Specifically, I utilized the AdamW optimizer [31] with a carefully tuned learning rate schedule incorporating warmup and decay, Layer-wise Learning Rate Decay (LLRD) [32], along with strong data augmentation, dropout [33], label smoothing [35], and Mixup [34].

The output of this stage is a ViT-L/14 visual encoder ($\mathbf{f}_{\theta'}$) whose weights θ' are primed with general medical visual knowledge, serving as the starting point for the subsequent stages of FusionMedCLIP.

3.2 Prompt Learning and Feature Alignment

Following the foundational adaptation in Stage 1, the ROCO-CLIP model, characterized by its visual encoder $\mathbf{f}_{\theta'}$ and text encoder $\mathbf{g}_{\phi}$, possesses a generalized understanding of medical visual-language concepts. However, adapting this foundation to specific downstream few-shot anomaly detection tasks requires further specialization. Fine-tuning the full visual backbone $\mathbf{f}_{\theta'}$ remains computationally demanding due to the large number of trainable parameters, and it also increases the risk of overfitting. [16]. Therefore, Stage 2 introduces a parameter-efficient adaptation strategy. This approach keeps the parameters of the pre-adapted visual encoder $\mathbf{f}_{\theta'}$ **frozen**, leveraging its powerful representations while avoiding prohibitive training costs. Instead, task-specific adaptation is achieved by training a set of **new, lightweight modules** built upon the extracted features: (1) a knowledge-augmented prompt learning component for the text encoder, and (2) a dedicated visual feature alignment Adapter for the visual encoder outputs. This strategy constitutes **transfer learning based on CLIP features, combined with prompt learning and visual feature alignment**, where the newly introduced modules collaboratively form the primary task predictor.

3.2.1 Knowledge-Augmented Prompt Learning

Manual prompt engineering for CLIP often yields suboptimal results, especially in specialized domains like medicine where nuanced descriptions are critical [12]. To create more effective and adaptable textual cues, I employ Context Optimization (CoOp) [11] for automated prompt generation. CoOp introduces a sequence of learnable continuous context vectors, $\{\mathbf{v}_1, \dots, \mathbf{v}_M\}$, prepended to the embedding of a class name (`{CLASS}`, e.g., 'normal', 'anomaly'). The combined sequence `"[v]1 ... [v]M {CLASS}"` is then input to the **frozen** CLIP text encoder $\mathbf{g}_\phi$. During Stage 2 training, only the context vectors $\mathcal{V}$ are updated, while the text encoder $\mathbf{g}_\phi$ remains frozen. However, especially with limited (few-shot) data, relying only on the simple class names 'normal' and 'anomaly' might not fully capture the required medical semantics. To **enrich** the learned prompts with established medical meaning, we **integrate knowledge** from the Unified Medical Language System (UMLS) [18]. Instead of just using the basic class name, this **CoOp+UMLS** approach leverages richer information associated with 'normal' and 'abnormal' concepts within the UMLS Metathesaurus. This includes medically relevant **synonyms** (e.g., 'pathological', 'WNL'), formal **definitions**, and related terms.

This structured medical knowledge is incorporated to **guide the learning** of the context vectors $\mathcal{V}$. Specifically, we leverage UMLS concepts pertinent to the classification task (e.g., 'normal' and 'abnormal'). Detailed examples of the UMLS information utilised, including Concept Unique Identifiers (CUIs), synonyms, and definitions, are provided in Appendix A **Table A3**. Our approach involves using this curated UMLS information, such as concept definitions or key synonyms (e.g., using "Pathological" or "Deviating from the normal condition" for 'abnormal'), to guide the initialisation of the learnable context vectors. We generate embeddings for these selected UMLS terms or definitions using CLIP text encorder (Transformer). These knowledge-derived embeddings are then used to initialise a subset of the context vectors $\{\mathbf{v}_m\}$ within $\mathcal{V}$. The overall process of assembling the context vectors (knowledge-initialised) and the class label is represented by the `ConstructPrompt` function within Eq. 3. Let $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_M\}$. The text encoder processes the constructed prompt to produce normalized, task-optimized text embeddings for the 'normal' ($\mathbf{f}_n$) and 'anomaly' ($\mathbf{f}_a$) classes:

$$\mathbf{f}_c = \mathrm{Normalize}\left(\mathbf{g}_\phi\left(\mathrm{ConstructPrompt}(\mathcal{V}, \mathrm{CLASS} = c)\right)\right), c \in \{\mathrm{normal}, \mathrm{anomaly}\} \quad (3)$$

3.2.2 Multi-Scale Visual Feature Extraction

Concurrent with prompt learning, the visual pathway utilizes the **frozen** pre-adapted visual encoder $\mathbf{f}_{\theta'}$ (ViT-L/14 architecture) to extract rich visual information. Drawing inspiration from the hierarchical nature of features in deep networks, where intermediate layers capture varying levels of abstraction potentially beneficial for dense prediction tasks [14], I extract feature maps from selected intermediate blocks. Specifically, I target the outputs of the 12th, 18th, and 24th blocks of the ViT-L/14 backbone. Let $\mathbf{G}_j(\mathbf{X}) \in \mathbb{R}^{H_j \times W_j \times C_j}$ denote the feature map from the j-th selected layer ($j \in \mathcal{J} = \{12,18,24\}$) for an input image **X**, where H_j, W_j are the spatial dimensions and C_j is the channel dimension.

3.2.3 Visual Feature Alignment via Lightweight Adapter

The raw visual features $\mathbf{G}_j(\mathbf{X})$ extracted from the frozen backbone $\mathbf{f}_{\theta'}$ are not inherently optimized for pixel-wise comparison with the learned text prompts $\mathbf{f}_n, \mathbf{f}_a$. Their channel dimensions C_j differ across layers and may not match the text embedding dimension C_{target} (e.g., 1024 for ViT-L/14). To bridge this modality and dimensionality gap, I introduce a trainable, lightweight visual feature alignment adapter. This adapter consists of a set of layer-specific modules, $\{\phi_j\}_{j \in \mathcal{J}}$, parameterized by learnable weights $\{\psi_j\}_{j \in \mathcal{J}}$. Each

ϕ_j is implemented as a simple **1x1 convolutional layer** that projects the input feature map $\mathbf{G}_j(\mathbf{X})$ from C_j channels to the target dimension C_{target}:

$$\mathbf{g}_j = \phi_j\big(\mathbf{G}_j(\mathbf{X}); \psi_j\big) \in \mathbb{R}^{H_j \times W_j \times C_{target}} \tag{4}$$

Crucially, the role of this adapter extends beyond mere dimensionality reduction; it aims to project the visual features into the *same semantic embedding space* as the textual features generated by CoOp+UMLS, thereby enabling meaningful similarity calculations. This projection preserves the spatial resolution (H_j, W_j) of the original feature map. Following the adapter, the feature vector at each spatial location (h, w) within $\mathbf{g}_j$ is L2-normalized to ensure consistent similarity scaling:

$$\tilde{\mathbf{g}}_j(h, w) = \frac{\mathbf{g}_j(h, w)}{\|\mathbf{g}_j(h, w)\|_2 + \epsilon} \tag{5}$$

where ϵ is a small constant for numerical stability. This adapter module serves as a parameter-efficient interface between the powerful but fixed visual backbone and the task-specific text prompts.

3.2.4 Similarity Mapping and Multi-Scale Aggregation

With the aligned, normalized visual features $\tilde{\mathbf{g}}_j(h, w)$ and the normalized text prompt embeddings $\mathbf{f}_n, \mathbf{f}_a$, I compute dense similarity maps reflecting the alignment of local image regions with 'normal' and 'anomaly' concepts. For each layer $j \in \mathcal{J}$ and spatial location (h, w), cosine similarities are calculated and transformed into probabilities using a Softmax function, scaled by the temperature parameter τ:

$$S_n^j(h, w) = \frac{\exp(\langle \tilde{\mathbf{g}}_j(h, w), \mathbf{f}_n \rangle / \tau)}{Z_j(h, w)}, \quad S_a^j(h, w) = \frac{\exp(\langle \tilde{\mathbf{g}}_j(h, w), \mathbf{f}_a \rangle / \tau)}{Z_j(h, w)} \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity (or dot product for normalized vectors), and the partition function $Z_j(h, w) = \exp(\langle \tilde{\mathbf{g}}_j(h, w), \mathbf{f}_n \rangle / \tau) + \exp(\langle \tilde{\mathbf{g}}_j(h, w), \mathbf{f}_a \rangle / \tau)$. This results in multi-scale probability maps $\mathbf{S}_n^j, \mathbf{S}_a^j \in \mathbb{R}^{H_j \times W_j}$.

To obtain a final prediction map at the original image resolution ($H \times W$), the anomaly probability maps from each scale are upsampled using bilinear interpolation and then averaged pixel-wise:

$$\mathbf{S}_a = \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} \text{Upsample}\left(\mathbf{S}_a^j, \text{size} = (H, W)\right) \in \mathbb{R}^{H\times W} \tag{7}$$

A corresponding normal probability map $\mathbf{S}_n$ can be computed similarly. The aggregated anomaly map $\mathbf{S}_a$ provides the final pixel-level localization result for this stage, which is used for calculating the training loss (detailed in Section 3.5).

3.3 End-to-End Lightweight Fine-tuning

While Stage 2 effectively adapts the frozen ROCOCLIP model using external modules, the visual features $\mathbf{G}_j(\mathbf{X})$ extracted by the frozen backbone $\mathbf{f}_{\theta'}$ may still not be optimally tailored for the specific nuances of the target medical anomaly detection task. To enable a deeper level of task-specific feature refinement without incurring the costs and risks of full fine-tuning, Stage 3 introduces **parameter-efficient, end-to-end lightweight fine-tuning** of the visual encoder, integrated seamlessly with the components developed in Stage 2.

3.3.1 Motivation for Deeper Visual Representation Adaptation

The core hypothesis motivating Stage 3 is that allowing targeted, minimal adjustments *within* the ViT-L/14 backbone can lead to visual representations (specifically at layers $j \in \mathcal{J} = \{12,18,24\}$) that are more discriminative for the downstream task of comparing against the learned CoOp+UMLS prompts $\mathbf{f}_n, \mathbf{f}_a$. By subtly modifying the internal feature transformations in a task-aware manner, this aim to enhance the model's ability to capture fine-grained details relevant to the specific anomalies, potentially improving detection and localization accuracy beyond what is achievable with fixed features alone, while strictly maintaining parameter efficiency suitable for few-shot learning.

3.3.2 Parameter-Efficient Fine-tuning Techniques

To achieve lightweight fine-tuning of the visual encoder $\mathbf{f}_{\theta'}$, I integrate two complementary parameter-efficient techniques, targeting different aspects of its computation while keeping the original parameters θ' frozen:

LoRA (Low-Rank Adaptation): I apply LoRA [19] to adapt the behaviour of key linear projection layers within *all* Transformer blocks of the ViT-L/14 backbone. Specifically, for the query (Q), key (K), and value (V) projection matrices in the multi-head self-attention mechanism, as well as the linear layers within the MLP blocks, I augment the original frozen weight matrix $\mathbf{W}_0$ (part of θ') with a low-rank update $\Delta\mathbf{W} = \mathbf{BA}$. Here, $\mathbf{A} \in \mathbb{R}^{r\times k}$ and $\mathbf{B} \in \mathbb{R}^{d\times r}$ are new, **trainable** low-rank matrices, with the rank r being significantly smaller than the original dimensions k and d ($r \ll \min(k, d)$). The forward pass through such a layer effectively becomes:

$$\mathbf{h} = \mathbf{W}_0\mathbf{x} + \mathbf{BAx} \tag{8}$$

Only the parameters of **A** and **B** are optimized during training, adding minimal parameter overhead. This allows for efficient adaptation of the internal feature transformations throughout the network depth.

Block Output Adapters: In addition to the layer-internal LoRA modifications, I introduce lightweight **Block Adapters** specifically at the output of the targeted intermediate Transformer blocks, namely the 12th, 18th, and 24th blocks ($j \in \mathcal{J}$). These adapters are small, **trainable** modules α_j, parameterized by ω_j, inserted directly after the respective block's computation (including layer normalization). It is important to note that this Block Adapterserves a distinct purpose focused on feature refinement, differing from the Alignment Adapter responsible for cross-modal preparation; this distinction will be further clarified in Section 3.3.3. Let $\mathbf{H}_j(\mathbf{X})$ be the output feature map of the j-th Transformer block *after* being influenced by internal LoRA modifications. The Block Adapter processes this output:

$$\mathbf{G}'_j(\mathbf{X}) = \alpha_j\big(\mathbf{H}_j(\mathbf{X}); \omega_j\big) \tag{9}$$

These adapters (e.g., bottleneck-style MLPs [29]) provide an additional locus for task-specific feature transformation, concentrated at the extraction points crucial for my downstream task, thus complementing the distributed adaptation provided by LoRA. The output $\mathbf{G}_j'(\mathbf{X})$ represents the *refined* visual feature map from layer j.

3.3.3 Integrated Architecture and End-to-End Training

In Stage 3, the data flow seamlessly integrates these new fine-tuning components with the Stage 2 architecture:

1. An input image **X** passes through the ViT-L/14 visual encoder ($\mathbf{f}_{\theta'}$). Its **base weights θ' remain frozen**.
2. Within each Transformer block, the QKV and MLP linear projections are modified by the **trainable LoRA** updates (as per Eq. 8).
3. At the output of blocks $j \in \mathcal{J} = \{12,18,24\}$, the resulting feature map $\mathbf{H}_j(\mathbf{X})$ passes through the corresponding **trainable Block Adapter** α_j (parameterized by ω_j) to yield refined feature maps $\mathbf{G}_j'(\mathbf{X})$ (Eq. 9).
4. These **refined** intermediate features $\mathbf{G}_j'(\mathbf{X})$ are then fed into the **Visual Feature Alignment Adapter** modules $\{\phi_j\}$ (parameterized by ψ_j) in Section 3.2.3. These adapters perform the same function as in Stage 2 – projecting features to dimension C_{target} – but now operate on potentially improved input features $\mathbf{G}_j'$. They are also **trainable** in this stage.

$$\mathbf{g}_j' = \phi_j\big(\mathbf{G}_j'(\mathbf{X}); \psi_j\big) \in \mathbb{R}^{H_j \times W_j \times C_{target}} \tag{10}$$

It is pertinent to clarify the distinct roles of the Block Adapter (α_j, Eq. 9) introduced in this stage and the Alignment Adapter (ϕ_j, Eq. 10) originating from Stage 2, despite their sequential application after specific ViT blocks ($j \in \mathcal{J}$). The **Block Adapter (α_j)** primarily focuses on **task-specific feature refinement**, acting as a lightweight, learnable module directly augmenting the internal representation produced by the j-th ViT block (influenced by LoRA) to better capture details relevant to the downstream anomaly detection objective. Conversely, the **Alignment Adapter (ϕ_j)**, implemented as a 1x1 convolution, serves a different critical function: **cross-modal alignment and**

dimension matching. Its purpose is to project the (potentially refined) visual features $\mathbf{G}'_j$ from possibly varying channel dimensions into the fixed CLIP embedding space (C_{target}), ensuring they reside in the same semantic space as the CoOp+UMLS text prompts ($\mathbf{f}_n, \mathbf{f}_a$) for meaningful cosine similarity comparison by the MapMaker, while preserving spatial resolution crucial for localization. Thus, α_j enhances the quality of visual features *within* the visual pathway based on task feedback, while ϕ_j prepares these features for effective *interaction* with the textual modality. They perform complementary, non-redundant functions essential to the overall FusionMedCLIP architecture. Following this alignment, L2-normalization is applied:

$$\tilde{\mathbf{g}}'_j(h,w) = \frac{\mathbf{g}'_j(h,w)}{\|\mathbf{g}'_j(h,w)\|_2 + \epsilon} \tag{11}$$

5. Concurrently, the CoOp+UMLS module generates the learnable text prompt embeddings $\mathbf{f}_n, \mathbf{f}_a$ using the **frozen** text encoder $\mathbf{g}_\phi$ and **trainable** context vectors $\mathcal{V}$, exactly as described by Eq.3 in Stage 2.
6. The similarity mapping and multi-scale aggregation (MapMaker) proceed identically to Stage 2's description (referencing Eq.6 and Eq.7), but now use the *refined* normalized visual features $\tilde{\mathbf{g}}'_j$ from Eq.12 and the same text prompts $\mathbf{f}_n, \mathbf{f}_a$ from Eq.6. Let $\mathbf{S}'_a$ denote the final anomaly probability map produced in this stage.

Critically, the training in Stage 3 is performed **end-to-end**. The optimization process updates *all* trainable parameters simultaneously using the composite loss function detailed in Section 3.5. The set of trainable parameters in Stage 3 comprises:

- The LoRA matrices $(\mathbf{A}, \mathbf{B})$ for all adapted layers in the ViT-L/14 visual encoder.
- The parameters ω_j of the Block Adapters $\{\alpha_j\}$ for $j \in \mathcal{J}$.
- The parameters ψ_j of the Visual Feature Alignment Adapters $\{\phi_j\}$ originally introduced in Stage 2.
- The CoOp+UMLS context vectors $\mathcal{V}$.
- The temperature parameter τ.

The original backbone parameters θ' of the visual encoder and ϕ of the text encoder (excluding prompt embeddings) remain **strictly frozen** throughout Stage 3 training. This end-to-end optimization allows the lightweight backbone modifications (LoRA, Block Adapter) and task-specific modules (CoOp, Alignment Adapter) to co-adapt synergistically.

3.4 Multi-method Anomaly Synthesis

Effective training of deep learning models for medical anomaly detection, especially under few-shot conditions, is often hampered by the scarcity of annotated abnormal samples [16]. To address this limitation and enhance the robustness of my model trained in Stages 2 and 3, I employ a multi-method anomaly synthesis strategy **exclusively during the training phase**. This involves generating synthetic anomaly images $\hat{\mathbf{X}}$ from readily available normal source images $\mathbf{X}$ using a corresponding anomaly mask $\mathbf{Y}$, represented generally as $\hat{\mathbf{X}} = \Psi(\mathbf{X}, \mathbf{Y})$. The goal is to augment the training dataset with a diverse set of synthesized anomalies that mimic various clinically relevant pathological patterns.

I integrate a suite of synthesis techniques, $\{\Psi_k\}$, chosen to represent different fundamental types of abnormalities encountered in medical imaging:

1. **Structural Alterations:** Methods simulating abrupt changes in structure, such as fractures, defects, or the presence of foreign objects. This includes:
   - *CutPaste-based methods:* Involving copying image patches and pasting them onto the target region $\mathbf{Y}$ [9].
   - *Enhanced Structural Synthesis:* This enhances realism by using seamless blending techniques. Often involves solving for the synthesized pixel values $\hat{\mathbf{X}}|_{\mathbf{Y}}$ within the mask $\mathbf{Y}$ such that their gradients $\nabla\hat{\mathbf{X}}$ closely match a target gradient field $\mathbf{V}$ (derived from a source patch or template), while ensuring boundary continuity with the surrounding region $\mathbf{X}|_{\partial\mathbf{Y}}$. This is commonly achieved via Poisson image editing principles.

$$\min_{\hat{\mathbf{X}}|_{\mathbf{Y}}} \iint_{\mathbf{Y}} ||\nabla\hat{\mathbf{X}} - \mathbf{V}||^2 d\mathbf{x} \text{ s.t. } \hat{\mathbf{X}}|_{\partial\mathbf{Y}} = \mathbf{X}|_{\partial\mathbf{Y}} \tag{12}$$

These methods aim to capture anomalies defined by sharp boundaries and structural discontinuity.

2. **Intensity and Density Variations:** Techniques simulating lesions primarily characterized by changes in pixel intensity or local density, representative of tumors, cysts, or inflammation. This includes:

    - *Gaussian Intensity Change:* Modifying pixel intensities within the mask $\mathbf{Y}$ based on filtered Gaussian noise $\tilde{\sigma}$: $\hat{\mathbf{X}}|_{\mathbf{Y}} = (\mathbf{X} + \gamma\tilde{\sigma})|_{\mathbf{Y}}$, where γ controls the intensity factor [36].

    - *Enhanced Density Synthesis:* This improves upon simple intensity changes by incorporating pathology-informed texture $\mathbf{T}_{tex}$ and boundary blurring. A blurred mask $\tilde{\mathbf{M}}_{\mathbf{Y}} = G_\sigma * \mathbf{M}_{\mathbf{Y}}$ (where G_σ is a Gaussian kernel informed by pathology profile) modulates the blending of the original image and the synthesized texture/intensity modification, governed by an intensity factor γ:

$$\hat{\mathbf{X}} = \mathbf{X} \odot \left(1 - \tilde{\mathbf{M}}_{\mathbf{Y}}\right) + (\mathbf{X} \cdot (1 - \beta) + \beta \cdot \mathbf{T}'_{tex}) \odot \tilde{\mathbf{M}}_{\mathbf{Y}} \qquad (13)$$

    where $\mathbf{T}'_{tex}$ could represent modified texture/intensity and β controls the mixing, or a similar additive/multiplicative model based on γ and $\mathbf{T}_{tex}$ is used. Inspired by methods like DRAEM [10].

3. **Deformation and Expansion:** Methods simulating tissue displacement, organ enlargement, or mass effects. This includes:

    - *Source-based Deformation:* Applying a warping field that pushes pixels within $\mathbf{Y}$ away from a central point c, transforming location l to $\hat{l}$ based on radius r and a deformation factor α: e.g., $\hat{l} = c + f(l, c, r, \alpha)$ [36].

    - *Guided Deformation:* This employs more sophisticated non-rigid transformations. The synthesis applies a warping function $\mathcal{T}_{\mathbf{u}}$ based on a displacement field $\mathbf{u}(\mathbf{p})$, such that $\hat{\mathbf{X}} = \mathbf{X} \circ \mathcal{T}_{\mathbf{u}}$, where $\mathcal{T}_{\mathbf{u}}(\mathbf{p}) = \mathbf{p} + \mathbf{u}(\mathbf{p})$ [Ref-ImageWarpingBook]. Crucially, the displacement field $\mathbf{u}(\mathbf{p})$ is generated considering anatomical constraints (e.g., defined centers c_{anat}, organ boundaries) and parameters derived from organ_profiles, aiming for physiologically plausible deformations rather than simple geometric expansions.

During training, for a given normal sample $\mathbf{X}$, one synthesis task Ψ_k from the combined pool (original CutPaste, GaussIntensityChange, Source, plus their enhanced counterparts) is randomly selected, along with a generated mask $\mathbf{Y}$. The resulting pair $(\hat{\mathbf{X}}, \mathbf{Y})$ is then used as input to the model for computing the loss (detailed in Section 3.5), with $\mathbf{Y}$ serving as the ground truth for segmentation objectives.

This multi-method synthesis strategy allows the model to learn from a wider variety of anomaly appearances than typically available in limited real datasets, fostering better generalization. I emphasize again that this synthesis is **strictly confined to the training stage** and is not applied during inference or evaluation, where the model processes original, unmodified query images.

3.5 Training Function and Inference Procedure

To effectively train the parameters of my model for both identifying the presence of anomalies (image-level detection) and pinpointing their exact location (pixel-level localization), I employ a composite loss function. This objective function combines losses tailored to each aspect of the task and allows for balancing their contributions via configurable weights. Let $\mathbf{S}'_a \in \mathbb{R}^{H \times W}$ be the final predicted anomaly probability map generated by the model (as described in Stage 3, Eq. 10, or $\mathbf{S}_a$ from Eq. 4 if referring to Stage 2). Let $\mathbf{Y} \in \{0,1\}^{H \times W}$ be the ground truth pixel-level anomaly mask (either from real annotations or synthesized during training as per Section 3.4), and let $y_{img} \in \{0,1\}$ be the corresponding image-level label indicating the presence (1) or absence (0) of any anomaly in the ground truth mask (i.e., $y_{img} = 1$ iff $\sum \ \mathbf{Y} > 0$).

The total loss L is computed as a weighted sum of an image-level classification loss L_{img} and a pixel-level segmentation loss L_{seg}:

$$L = w_{img} L_{img} + w_{seg} L_{seg} \tag{14}$$

- **Image-Level Loss (L_{img}):** To address the classification aspect and handle potential class imbalance between normal and abnormal images, I utilize the Focal Loss. It is computed between image-level prediction derived from $\mathbf{S}'_a$ (e.g., using the maximum predicted probability, $s_{pred} = \max(\mathbf{S}'_a)$) and the ground truth image label y_{img}.

$$L_{img} = \text{FocalLoss}\left(s_{pred}, y_{img}\right) \tag{15}$$

- **Pixel-Level Segmentation Loss (L_{seg}):** To optimize for accurate spatial localization of anomalies, I employ the Dice Loss, specifically its binary variant, which measures the overlap between the predicted anomaly map $\mathbf{S}'_a$ and the ground truth mask **Y** (after ensuring **Y** is appropriately thresholded/binarized if needed).

$$L_{seg} = \text{BinaryDiceLoss}(\mathbf{S}'_a, \mathbf{Y}) \tag{16}$$

The weights w_{img} and w_{seg} are hyperparameters defined in the configuration, allowing adjustment of the relative importance of overall detection versus precise localization during training. The gradients derived from this total loss L are used to update all trainable parameters defined for the respective training stage (described in Sections 3.2.4 and 3.3.3).

Once the model has been trained (typically using the end-to-end regime of Stage 3), the inference phase applies the learned model to unseen query images $\mathbf{X}_{query}$ to generate anomaly detection and localization predictions. Importantly, the anomaly synthesis techniques described in Section 3.4 are **not** used during inference.

The procedure for a given query image $\mathbf{X}_{query}$ is as follows:

1. **Forward Pass:** The image $\mathbf{X}_{query}$ is fed into the trained visual encoder $\mathbf{f}_{\theta'}$. If the model was trained through Stage 3, the embedded LoRA modules (Eq. 8) and Block Adapters (Eq. 9) modify the feature computation using their learned parameters. The base weights θ' remain frozen.
2. **Feature Extraction & Refinement:** Refined multi-scale visual features $\mathbf{G}'_j(\mathbf{X}_{query})$ are extracted from layers $j \in \mathcal{J}$.

3. **Visual Alignment:** These features are processed by the trained Visual Feature Alignment Adapters $\{\phi_j\}$ (Eq. 10) and normalized (Eq. 11) to produce $\tilde{\mathbf{g}}'_j(h, w)$.
4. **Prompt Application:** The fixed, optimized CoOp+UMLS text prompt embeddings $\mathbf{f}_n$ and $\mathbf{f}_a$ (from Eq. 3) are used.
5. **Similarity Mapping & Aggregation:** Pixel-wise similarities are computed, converted to probabilities (using the learned τ), and aggregated across scales to yield the final anomaly probability map $\mathbf{S}'_a \in \mathbb{R}^{H\times W}$.
6. **Output Generation:**
   - **Pixel-Level Localization:** The resulting anomaly map $\mathbf{S}'_a$ directly serves as the pixel-level prediction, indicating the probability of anomaly at each location. It can be thresholded for visualization or quantitative segmentation evaluation.
   - **Image-Level Detection:** An image-level anomaly score s_{img} is derived from the pixel map $\mathbf{S}'_a$. A common practice, which I adopt, is to use the maximum probability value within the map: $s_{img} = \max_{(h,w)} \mathbf{S}'_a(h, w)$. This score s_{img} is then used for image classification metrics (e.g., AUC, Accuracy) by comparing against a threshold.

This inference pipeline leverages the synergy between the (potentially fine-tuned) visual features and the optimized textual prompts to perform zero-shot or few-shot anomaly assessment on new medical images.

Chapter 4

Experiment and Results

This section details the comprehensive experimental evaluation conducted to assess the efficacy of the proposed FusionMedCLIP framework. My primary objective is to demonstrate its performance advantages in few-shot medical image anomaly detection and localization across diverse datasets and modalities. I present comparative results against representative baseline methods, rigorously analyze the contribution of key components through ablation studies and investigate the impact of varying few-shot sample sizes.

4.1 Dataset and Evaluation Metrics

Dataset: To rigorously evaluate the proposed FusionMedCLIP framework, I conducted comprehensive experiments focusing on its effectiveness in few-shot medical image anomaly detection and localization. My evaluation spans three diverse public datasets: **CheXpert** [22] (Chest X-ray, various pathologies), **Brain MRI** [20] (MRI Tumor Detection, brain tumors), and **BUSI** [21] (Breast Ultrasound for benign/malignant classification, treated as anomalies). I specifically target the challenging few-shot scenario where only $k \in \{4,8,16,32\}$ normal images per dataset are available for model training and adaptation. Crucially, my training paradigm leverages these limited normal samples by applying the multi-method anomaly synthesis techniques (Section 3.4) to generate training pairs, enabling the optimization of FusionMedCLIP's learnable components (CoOp prompts, Adapters, LoRA). Standard preprocessing was applied, resizing images to 224×224 and normalizing using CLIP's defaults [1].

Dataset	*Modality*	*Anomaly Examples*	*Test Set Size (Normal/Abnormal)*	*Few-Shot Training (k Normal Samples)*
CheXpert	Chest X-ray	Cardiomegaly, Edema, etc.	250 / 250	$K \in \{4,8,16,32\}$
Brain MRI	Brain 2D mri	Tumors	65 / 155	
BUSI	Ultrasound	Benign & Malignant Tumors	101 / 647	

Table 5. Overview of Datasets used for Evaluation

Evaluation Metrics: I assess performance using standard metrics appropriate for anomaly detection tasks. **Image-level** anomaly detection performance is measured by the Area Under the Receiver Operating Characteristic Curve (**Image-AUROC**), quantifying the model's ability to distinguish between normal and abnormal images based on the maximum predicted anomaly score. **Pixel-level** anomaly localization accuracy is evaluated using the **Pixel-AUROC**, computed across all pixels in the test set, reflecting the model's capability to accurately delineate anomalous regions within an image via the predicted map $\mathbf{S}'_{a}$.

4.2 Implementation Details

Detailed hyperparameters for the FusionMedCLIP training are provided in Appendix A (Table A1). FusionMedCLIP framework builds upon a ViT-L/14 model pre-trained using CLIP objectives and subsequently fully fine-tuned on the ROCO dataset [17] **(training details in Appendix A, Table A2)**, serving as frozen backbone (ROCOCLIP). The core architecture utilizes features from the 14th, 18th, and 24th layers of this backbone. Key learnable components include **CoOp+UMLS** prompts ($M = 8$ learnable tokens), a **1x1 Convolution-based Alignment Adapter** mapping features to the CLIP embedding dimension (1024), and **Block Adapters** with a bottleneck dimension of 64 inserted after the target layers. Parameter-efficient fine-tuning was enabled via **LoRA** (rank $r = 8$, alpha $\alpha = 16$) applied to the Q, K, V, output projections and MLP layers within all ViT blocks.

The model was trained end-to-end using the **AdamW optimizer** [31] for 200 epochs with a batch size of 8. The composite loss function (Eq. 1) equally weighted the image-level (Focal Loss) and pixel-level (Dice Loss) contributions ($w_{img} = 1.0, w_{seg} = 1.0$). Anomaly synthesis tasks (Section 3.4), were applied with equal probability during training. Two methods were employed to generate binary anomaly mask Y. Varied anomaly shapes were initially created using a Perlin noise generator, with the output subsequently binarised. Alternatively, for tasks needing the radius of the anomalous region to be calculated, geometric shapes such as ellipses or rectangles, featuring randomised sizes and rotation angles, served as anomaly masks. All experiments were repeated three times with different random seeds and sampled k-shot normal support sets; average results are reported.

4.3 Baselines and Results

For a comprehensive assessment of FusionMedCLIP's few-shot AD capabilities, a varied selection of baseline methods was established for comparative analysis. These include methods based on deep feature embedding (PatchCore [37], SimpleNet [38]), image reconstruction (SQUID [39], AnoDDPM [40]), self-supervised learning (CutPaste [9]), and a dedicated few-shot anomaly detection approach (RegAD [41]). Crucially, for fair comparison, all external baselines were re-implemented under specific few-shot setting ($k \in \{4,8,16,32\}$ normal training images) on the CheXpert, BrainMRI, and BUSI datasets, utilizing the same anomaly synthesis strategy (Section 3.4) during their training phase where applicable and beneficial. Performance was evaluated using **Image-AUROC** for anomaly detection and **Pixel-AUROC** for anomaly localization, with higher values indicating better performance. All experiments were repeated 3 times with results averaged.

Dataset	*k*	*PatchCore*	*SimpleNet*	*SQUID*	*AnoDDPM*	*CutPaste*	*RegAD*	***FusionMedCLIP***
CheXpert	4	60.0±1.8	63.5±2.4	66.0±1.9	55.8±1.3	63.6±2.8	59.8±1.4	**71.1±1.8**
	8	59.9±2.2	63.8±0.7	66.6±3.9	57.3±3.5	65.0±2.9	62.0±3.0	**71.8±1.3**
	16	64.6±1.6	65.6±3.7	67.6±4.7	59.7±2.8	65.2±3.2	60.2±1.7	**73.6±1.4**
	32	64.3±1.2	67.6±1.1	69.4±2.2	66.9±3.2	71.4±2.0	65.6±1.4	**74.3±1.1**
BrainMRI	4	70.8±3.5	77.0±3.0	69.5±4.4	67.2±1.8	72.3±1.8	67.5±1.0	**93.5±0.6**
	8	75.9±2.9	79.3±2.9	75.8±1.9	73.0±3.6	77.0±4.9	75.1±2.8	**94.0±0.9**
	16	79.5±1.9	81.7±4.4	77.2±0.6	77.2±2.5	80.2±2.7	82.4±1.6	**94.8±0.8**
	32	81.7±0.2	82.6±0.9	79.3±1.7	80.0±2.1	79.0±4.3	83.0±2.4	**95.4±0.2**
BUSI	4	81.7±1.3	76.1±2.5	55.4±2.2	69.3±1.3	75.7±1.4	74.7±5.0	**88.8±1.3**
	8	82.6±2.1	80.4±4.8	58.7±1.6	72.6±4.6	78.2±2.1	75.8±2.7	**89.6±0.9**
	16	87.6±1.0	84.1±2.7	64.8±4.5	74.4±2.9	80.0±0.9	76.5±1.2	**91.5±0.4**
	32	88.9±0.1	88.0±0.3	67.8±0.4	76.2±4.0	80.2±1.7	78.6±1.2	**92.1±1.0**

Table 6. Comparing AD Performance Under Few-shot Conditions: FusionMedCLIP relative to other methods, evaluated with Image-AUROC (%).

k	*PatchCore*	*SimpleNet*	*SQUID*	*AnoDDPM*	*RegAD*	***FusionMedCLIP***
4	78.5±2.6	73.5±4.9	56.8±5.1	68.3±1.9	71.9±3.1	**88.8±1.3**
8	78.1±1.3	77.6±3.4	56.8±1.2	74.7±4.3	76.1±2.6	**89.6±0.9**
16	80.2±0.8	78.2±1.5	65.5±2.8	79.5±1.4	74.6±1.0	**91.5±0.4**
32	80.3±0.1	79.8±0.4	68.4±2.2	79.1±1.3	76.3±0.4	**92.1±1.0**

Table 7. Evaluating Anomaly Localization on the BUSI Dataset: FusionMedCLIP's performance benchmarked against other methods using Pixel-AUROC (%).

FusionMedCLIP consistently demonstrates superior anomaly detection across all datasets and few-shot settings (**Table 6**). Notably, on BrainMRI (k=4), it achieves 93.5% Image-AUROC, significantly surpassing the best external baseline (SimpleNet, 77.0%) by >16%. Similarly, on BUSI (k=4), it outperforms PatchCore (81.7%) with 88.8%. Even on CheXpert, FusionMedCLIP maintains a clear advantage (e.g., 71.1% vs 66.0% at k=4). This robust performance, especially in critical low-shot regimes, highlights the effectiveness of our CLIP adaptation strategy over traditional methods for leveraging pre-trained knowledge in data-scarce medical scenarios.

The anomaly localization results on BUSI (**Table 7**) further confirm FusionMedCLIP's superiority. Our method achieves significantly higher Pixel-AUROC, reaching 88.8% even at k=4 (vs. 78.5% for PatchCore, >10% gain). This advantage persists across all k values, peaking at 92.1% for k=32. This demonstrates FusionMedCLIP's strong capability for both detecting and precisely localizing anomalies, a crucial aspect for clinical utility, validating our fused approach for detailed spatial understanding.

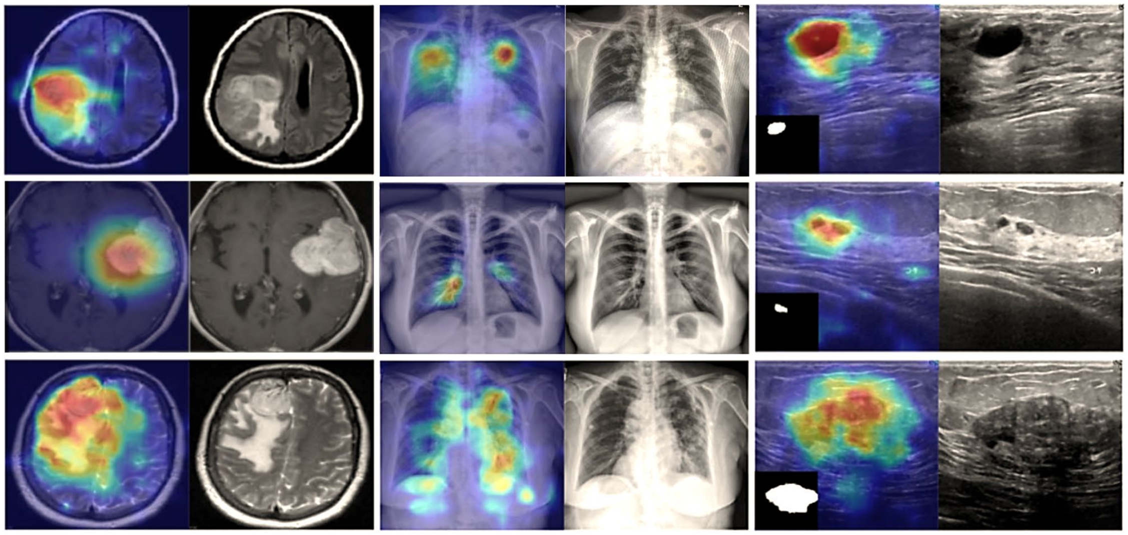

Figure 4. Qualitative Evaluation of FusionMedCLIP: showing the anomaly image and its predicted anomaly map for each group.

4.4 Ablation Studies

To meticulously evaluate the individual contribution of each core innovation integrated into FusionMedCLIP, we conducted a progressive ablation study. Starting from the basic CLIP architecture, we incrementally introduced our proposed components: Stage 1 domain pre-training (ROCOCLIP), learnable prompts (CoOp and CoOp+UMLS), the 1x1 Convolution-based Alignment Adapter (ϕ_j), and finally the Stage 3 lightweight fine-tuning (LoRA + Block Adapters α_j). This additive approach allows for a clear assessment of the performance gain attributable to each element. All ablation experiments were performed under the $k = 16$ few-shot setting using anomaly synthesis for training. When learnable prompts (CoOp/CoOp+UMLS) were not used, we employed standard manual prompts (e.g., "a photo of a {}"). When the Alignment Adapter (ϕ_j) was absent, we utilized global average pooling (GAP) over the spatial dimensions of the extracted ViT features before computing similarity with text embeddings.

The following configurations were evaluated progressively:

1. **CLIP ViT-L/14 (Baseline):** Original OpenAI model + Manual Prompts + GAP.
2. **+ ROCOCLIP:** Adds Stage 1 Pre-training. (ROCOCLIP + Manual Prompts + GAP).
3. **+ CoOp:** Adds basic Prompt Learning. (ROCOCLIP + CoOp + GAP).
4. **+ UMLS:** Adds UMLS knowledge. (ROCOCLIP + CoOp + UMLS + GAP).
5. **+ Align Adapter (ϕ_j):** Adds specialized feature alignment (replaces GAP). (ROCOCLIP + CoOp+UMLS + Align Adpt ϕ_j)

	Configuration	*CheXpert*	*BrainMRI*	*BUSI*	*Key Added Component / Change*
1	CLIP ViT-L/14	41.6±1.3	40.5±0.1	49.5±0.6	Baseline
2	+ ROCOCLIP	53.3±0.7	57.4±1.3	66.8±0.1	+ Stage 1 Pre-training
3	+ CoOp	60.3±2.1	73.8±0.3	81.1±1.1	+ Basic Learnable Prompts
4	+ UMLS	62.7±1.3	77.5±1.4	83.6±0.7	+ UMLS Knowledge Enhancement
5	+ Align Adapter (ϕ_j)	71.8±1.2	93.3±0.5	90.3±0.3	+ Alignment Adapter (Replaces GAP)
6	FusionMedCLIP	**73.6±1.4**	**94.8±0.8**	**91.5±0.4**	**+ LoRA & Block Adapter FT**

Table 8. Progressive ablation study under k=16 setting (Image-AUROC %)

6. **+ FusionMedCLIP:** Adds LoRA & Block Adapters fine-tuning. (ROCOCLIP + CoOp+UMLS + Align Adpt ϕ_j + LoRA + Block Adpt α_j)

The results of the progressive ablation study, presented in **Table 8**, clearly demonstrate the effectiveness of each subsequent component added to the FusionMedCLIP framework. Starting from the basic zero-shot performance of the original CLIP model with manual prompts and GAP feature aggregation (Row 1), applying Stage 1 ROCO pre-training (Row 2) provides a **substantial initial boost** across all datasets (e.g., +11.7% on CheXpert, +16.9% on BrainMRI, +17.3% on BUSI Image-AUROC), signifying the advantage of domain adaptation. Replacing manual prompts with learnable CoOp prompts (Row 3) further improves performance considerably (e.g., +7.0% on CheXpert, +16.4% on BrainMRI, +14.3% on BUSI vs Row 2), indicating the power of automated prompt tuning even with simple feature aggregation. Enhancing CoOp with UMLS knowledge (Row 4) yields a **modest yet consistent additional gain** (approx. +2.4-3.7% vs Row 3), confirming the value of external semantic guidance. A **significant leap in performance** is observed upon introducing the 1x1 Convolution-based Alignment Adapter which replaces GAP (Row 5 vs Row 4, e.g., +9.1% on CheXpert, +15.9% on BrainMRI, +6.7% on BUSI), highlighting the critical importance of preserving spatial information and properly aligning visual features for dense comparison; this configuration represents our full Stage 2 model. Finally, the incorporation of Stage 3's lightweight end-to-end fine-tuning through LoRA and Block Adapters (Row 6, Full FusionMedCLIP) delivers the **highest performance across all metrics** (e.g., average +1.4% Image-AUROC gain over Row 5), solidifying the benefit of synergistically optimizing both the learned prompts/adapters and the visual feature extraction process itself. This step-by-step improvement robustly validates the rationale behind our multi-component fusion approach.

Chapter 5
Conclusion and Future Work

5.1 Conclusion

In this dissertation, we addressed the significant challenge of few-shot anomaly detection and localization in medical imaging by proposing **FusionMedCLIP**, a multi-stage framework designed to effectively adapt the powerful vision-language model CLIP. Our approach synergistically integrates (1) initial domain adaptation via fine-tuning on a broad medical dataset (ROCO), (2.1) task-specific adaptation using knowledge-augmented learnable prompts (CoOp+UMLS) and (2.2) a dedicated 1x1 convolutional Alignment Adapter (ϕ_j) operating on frozen CLIP features, and (3) further performance enhancement through parameter-efficient, end-to-end fine-tuning of the visual backbone using LoRA and Block Adapters (α_j). Critically, our training methodology leverages multi-method anomaly synthesis to overcome the inherent data scarcity of the few-shot setting.

Comprehensive experiments conducted across diverse medical imaging datasets (CheXpert, Brain MRI, BUSI) and modalities demonstrated the effectiveness of FusionMedCLIP. Our framework consistently outperformed a wide range of competitive external baselines in both image-level anomaly detection (Image-AUROC) and pixel-level anomaly localization (Pixel-AUROC), particularly under stringent few-shot conditions. Furthermore, our progressive ablation studies rigorously validated the contributions of each core component: the initial ROCO pre-training provided essential domain grounding, the combination of CoOp+UMLS and the Alignment Adapter established a strong baseline for few-shot adaptation, and the subsequent end-to-end lightweight fine-tuning (Stage 3) delivered substantial performance gains, highlighting the benefits of optimizing visual features and prompt representations concurrently. The necessity of anomaly synthesis for effective training in this low-data regime was also confirmed.

5.2 Future Work

Despite the promising results, this work has certain limitations. The performance of FusionMedCLIP, particularly its reliance on anomaly synthesis during training, may be sensitive to the realism and diversity of the synthesis methods employed. While our enhanced synthesis tasks aim for clinical relevance, bridging the gap between synthetic and real, potentially subtle anomalies remains an ongoing challenge. Additionally, the interpretability of the learned prompts and the internal workings of the fine-tuned CLIP model warrant further investigation. The computational overhead introduced by the end-to-end fine-tuning (Stage 3), although parameter-efficient compared to full fine-tuning, is higher than the frozen-backbone approach (Stage 2).

Future research directions stemming from this work are manifold. Exploring more sophisticated and pathology-specific anomaly synthesis techniques, possibly leveraging generative adversarial networks (GANs) or advanced diffusion models conditioned on clinical knowledge, could further enhance training robustness. Investigating the framework's applicability to a wider range of medical imaging modalities (e.g., CT, Pathology slides) and extending it towards multi-class few-shot anomaly classification or direct segmentation are promising avenues. Incorporating techniques to improve model interpretability, such as visualizing attention maps or analyzing learned prompt embeddings, would increase clinical trust. Finally, conducting prospective clinical validation studies to assess the real-world utility and generalizability of FusionMedCLIP is an essential next step towards potential clinical translation.

In conclusion, FusionMedCLIP offers a robust and effective strategy for adapting large vision-language models to the critical task of few-shot medical image anomaly detection and localization. By combining domain pre-training, advanced prompt learning, feature alignment, and parameter-efficient fine-tuning, the framework provides a significant step forward in leveraging foundational models for data-scarce medical imaging applications.

References

[1] Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision,” *International Conference on Machine Learning*, pp. 8748–8763, Feb. 2021.

[2] D. Gong *et al.*, “Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019. doi: 10.1109/iccv.2019.00179.

[3] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, “MedCLIP: Contrastive Learning from Unpaired Medical Images and Text,” Conference on Empirical Methods in Natural Language Processing.

[4] H. Zhao, H. Li, and L. Cheng, “Synthesizing Filamentary Structured Images with GANs,” *arXiv.org*, vol. abs/1706.02185, Jun. 2017.

[5] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning,” *Nat. Biomed. Eng*, vol. 6, no. 12, pp. 1399–1406, Sep. 2022, doi: 10.1038/s41551-022-00936-9.

[6] X. Chen, Y. He, C. Xue, R. Ge, S. Li, and G. Yang, “Knowledge Boosting: Rethinking Medical Contrastive Vision-Language Pre-Training,” International Conference on Medical Image Computing and Computer-Assisted Intervention.

[7] K. You *et al.*, “CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training,” in *Lecture Notes in Computer Science*, Cham: Springer Nature Switzerland, 2023, pp. 101–111. doi: 10.1007/978-3-031-43895-0_10.

[8] H. M. Schlüter, J. Tan, B. Hou, and B. Kainz, “Natural Synthetic Anomalies for Self-supervised Anomaly Detection and Localization,” in *European Conference on Computer Vision*, Cham: Springer Nature Switzerland, 2022, pp. 474–489. doi: 10.1007/978-3-031-19821-2_27.

[9] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, “CutPaste: Self-Supervised Learning for Anomaly Detection and Localization,” in *Computer Vision and Pattern Recognition*, IEEE, Jun. 2021, pp. 9659–9669. doi: 10.1109/cvpr46437.2021.00954.

[10] V. Zavrtanik, M. Kristan, and D. Skocaj, “DRÆM – A discriminatively trained reconstruction embedding for surface anomaly detection,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2021, pp. 8310–8319. doi: 10.1109/iccv48922.2021.00822.

[11] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to Prompt for Vision-Language Models,” *Int J Comput Vis (IJCV)*, vol. 130, no. 9, pp. 2337–2348, Jul. 2022, doi: 10.1007/s11263-022-01653-1.

[12] Y. Cai, H. Chen, and K.-T. Cheng, “Rethinking Autoencoders for Medical Anomaly Detection from A Theoretical Perspective,” International Conference on Medical Image Computing and Computer-Assisted Intervention.

[13] X. Zhang, M. Xu, D. Qiu, R. Yan, N. Lang, and X. Zhou, “MediCLIP: Adapting CLIP for Few-shot Medical Image Anomaly Detection,” International Conference on Medical Image Computing and Computer-Assisted Intervention.

[14] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *International Conference on Learning Representations*, vol. abs/2010.11929, Oct. 2020.

[15] W. F. Shah, “Preserving Privacy and Security: A Comparative Study of Health Data Regulations - GDPR vs. HIPAA,” *IJRASET*, vol. 11, no. 8, pp. 2189–2199, Aug. 2023, doi: 10.22214/ijraset.2023.55551.

[16] S. Woerner and C. F. Baumgartner, “Navigating Data Scarcity using Foundation Models: A Benchmark of Few-Shot and Zero-Shot Learning Approaches in Medical Imaging”, *MedAGI@MICCAI*, Aug. 15, 2024.

[17] O. Pelka, S. T. M. Rukavina, D. Shapira, K. Morik, and J. M. T. Koitka, "ROCO: Radiology Objects in COntext, a Multimodal Image-Caption Dataset," *arXiv preprint arXiv:2107.04609*, 2021.

[18] O. Bodenreider, "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology," *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D267–D270, 2004.

[19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[20] B. H. Menze et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, 2015.

[21] W. Al-Dhabyani, M. Gomaa, H. Khaled, and F. Fahmy, "Dataset of Breast Ultrasound Images," *Data Brief*, vol. 28, p. 104863, 2020.

[22] J. Irvin *et al.*, “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison”, *AAAI*, vol. 33, no. 01, pp. 590–597, Jul. 2019, doi: 10.1609/aaai.v33i01.3301590.

[23] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, “Deep Metric Learning via Lifted Structured Feature Embedding”, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 4004–4012. doi: 10.1109/cvpr.2016.434.

[24] W. Song, D. Wu, W. Shen, and B. Boulet, “Meta-Learning Based Early Fault Detection for Rolling Bearings via Few-Shot Anomaly Detection”, *arXiv.org*, Jan. 01, 2022.

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer, 2015, pp. 234–241.

[27] J. Snell, K. Swersky, and R. Zemel, “Prototypical Networks for Few-shot Learning,” *Neural Information Processing Systems*, pp. 4077–4087, Mar. 2017.

[28] C. Finn, P. Abbeel, and S. Levine, “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks,” *International Conference on Machine Learning*, pp. 1126–1135, Mar. 2017.

[29] P. Gao *et al.*, “CLIP-Adapter: Better Vision-Language Models with Feature Adapters,” *Int J Comput Vis*, vol. 132, no. 2, pp. 581–595, Sep. 2023.

[30] J. Lee *et al.*, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”, *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Sep. 2019, doi: 10.1093/bioinformatics/btz682.

[31] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[32] J. Howard and S. Ruder, “Universal Language Model Fine-tuning for Text Classification,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 328–339. doi: 10.18653/v1/p18-1031.

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[34] H. Zhang, M. Cissé, Y. Dauphin, and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” *International Conference on Learning Representations*, vol. abs/1710.09412, Oct. 2017.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[36] M. Baugh, J. Tan, J. P. Müller, M. Dombrowski, J. Batten, and B. Kainz, "Many Tasks Make Light Work: Learning to Localise Medical Anomalies from Multiple Synthetic Tasks," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 14220. Cham, Switzerland: Springer, 2023, pp. 172–182, doi: 10.1007/978-3-031-43907-0_16.

[37] K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, “Towards Total Recall in Industrial Anomaly Detection,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022. doi: 10.1109/cvpr52688.2022.01392.

[38] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, “SimpleNet: A Simple Network for Image Anomaly Detection and Localization,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2023, pp. 20402–20411. doi: 10.1109/cvpr52729.2023.01954.

[39] T. Xiang *et al.*, “SQUID: Deep Feature In-Painting for Unsupervised Anomaly Detection,” in *Computer Vision and Pattern Recognition*, IEEE, Jun. 2023, pp. 23890–23901. doi: 10.1109/cvpr52729.2023.02288.

[40] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, “AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2022, pp. 649–655. doi: 10.1109/cvprw56347.2022.00080.

[41] C. Huang, H. Guan, A. Jiang, Y. Zhang, M. Spratling, and Y.-F. Wang, “Registration based Few-Shot Anomaly Detection”, *European Conference on Computer Vision*, Jan. 01, 2022. doi: 10.48550/arxiv.2207.07361.

Appendix A. Implementation Parameter

This section provides detailed parameters used in our experiments. Table A1 lists the settings for training the main FusionMedCLIP framework (Stages 2 and 3). Table A2 details the configuration used for the Stage 1 fine-tuning of the ViT-L/14 model on the ROCO dataset, which yielded the ROCOCLIP backbone used in subsequent stages. Table A3 provides a summary of the Unified Medical Language System (UMLS) information utilised concepts within FusionMedCLIP as described in Section 3.2.1

PARAMETER	**VALUE**
BASE MODEL	ViT-L/14 (ROCOCLIP backbone)
TARGET VIT LAYERS ($\mathcal{J}$)	14, 18, 24
IMAGE SIZE	224 × 224
EPOCHS	200
BATCH SIZE	8
OPTIMIZER	AdamW
WEIGHT DECAY	0.01
OPTIMIZER BETAS	(0.9, 0.999)
OPTIMIZER EPSILON	1e-8
LOSS WEIGHTS (W_{IMG}/W_{SEG})	1.0 / 1.0
COOP+UMLS SETTINGS	
COOP LEARNABLE TOKENS (M)	8
COOP CSC	True
COOP TOKEN POSITION	end
COOP LR	1e-3
ALIGNMENT ADAPTER (Φ_J) SETTINGS	
ADAPTER STRUCTURE	1x1 Convolution (Layer Input Channels → 1024)
ALIGNMENT ADAPTER LR	1e-3
LORA SETTINGS	
LORA RANK (R)	8
LORA ALPHA (A)	16
LORA DROPOUT	0.05
LORA TARGET MODULES	q_proj, k_proj, v_proj, out_proj, fc1, fc2
LORA LR	3e-5
BLOCK ADAPTER (A_J) SETTINGS	
ADAPTER BOTTLENECK SIZE	64
BLOCK ADAPTER DROPOUT	0.1
BLOCK ADAPTER LR	5e-5
OTHER SETTINGS	
TEMPERATURE (T)	0.07 (fixed)
ANOMALY SYNTHESIS PROB.	Equal for Identity, Enhanced Structural, Density, Deformation

Table A1: Hyperparameters for FusionMedCLIP Training (Stages 2 & 3).

PARAMETER	VALUE
BASE MODEL	ViT-L/14 (CLIP pre-trained)
DATASET	ROCO
MAX. EPOCHS	50
BATCH SIZE	32
INPUT SIZE	(224, 224)
OPTIMIZER	AdamW
BASE LR	3e-6
WEIGHT DECAY	0.05
OPTIMIZER BETAS	(0.9, 0.999)
OPTIMIZER EPS	1e-8
LR SCHEDULER	warmup_cosine
WARMUP EPOCHS	5
MIN LR	1e-7
LAYER-WISE LR DECAY	True
LAYER DECAY RATE	0.8
LABEL SMOOTHING	0.1
MIXUP ALPHA	0.2
TEXT MIXUP	True
DROPOUT RATE	0.1
EARLY STOPPING PATIENCE	5
EARLY STOPPING MIN DELTA	1e-4
TRANSFORMS (TRAIN)	RandomResizedCrop, HFlip, ColorJitter, Rotation, Grayscale
TRANSFORMS (TEST)	ShorterResizeCrop, CenterCrop
MAX SEQ LENGTH	77

Table A2: Hyperparameters for Stage 1 ROCOCLIP Fine-tuning on ROCO.

	Normal	***Abnormal***
Concept Identifier	**C0231218**	**C0000768**
Synonyms & Variants	Within normal limits	Pathological
	Unremarkable	Irregular
	Negative	Deviant
	WNL	Diseased
	Physiological	Unusual
	Healthy	Atypical
	Normal appearance	Aberrant
	No significant findings	Anomalous
	No abnormality detected	Not within normal limits
	Negative findings	Positive findings
Definitions	A state or condition that is considered standard or typical.	A state indicating the presence of disease or pathology.
	The absence of disease or abnormality.	Deviating from the normal or expected condition.
	Findings within expected limits.	Findings that are not typical or standard
Semantic Types	Finding / Qualitative Concept / Pathologic Function	

Table A3: UMLS Information for 'Normal' and 'Abnormal' Concepts.