Integrating Vision–Language Models with Medical Imaging and Clinical Data for Improved Lymphoma Diagnosis

Master's thesis in Electrical Engineering

Yihan Tang

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Integrating Vision–Language Models with Medical Imaging and Clinical Data for Improved Lymphoma Diagnosis

Yihan Tang

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Division of Signal Processing and Biomedical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2026

Integrating Vision–Language Models with Medical Imaging and Clinical Data for Improved Lymphoma Diagnosis
Yihan Tang

Supervisor: Ida Häggström, Department of Electrical Engineering
Examiner: Ida Häggström, Department of Electrical Engineering

Master's Thesis 2026
Department of Electrical Engineering
Division of Signal Processing and Biomedical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Integrating Vision–Language Models with Medical Imaging and Clinical Data for Improved Lymphoma Diagnosis

Yihan Tang
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Recent advancements in deep learning have fundamentally transformed the field of medical imaging. Along with the increasing burden of cancer worldwide, the demand for automated diagnostic tools has surged. An example of such process is the Lymphoma Artificial Reader System (LARS), an ensemble model composed of 10 ResNet34-based models trained on over 17,000 [^{18}F] FDG-PET/CT scans. Although LARS has achieved impressive diagnostic accuracy, it relies solely on image features and lacks the capability to understand and utilize patient metadata (such as age, sex, and smoking history) and treatment information. However, these structured clinical records play a crucial role in diagnostic decisions in daily clinical practice.
In order to bridge this gap, we propose a new vision–language version of LARS (namely MM-LARS) capable of incorporating patient context via a tabular-to-text transformation. By converting structured patient data into natural language prompts, we effectively fuse clinical text with dual-view PET images. Through extensive ablation experiments, we demonstrate that this multimodal approach not only improves overall diagnostic performance over vision-only baselines, but also significantly boosts model specificity. Notably, the inclusion of clinical text successfully corrects false-positive misdiagnoses caused by ambiguous visual features. Our work provides practical insight into the next generation of multimodal diagnostic AI, offering a framework that is potentially generalizable beyond lymphoma.

Keywords: Medical Imaging, Vision-language Models, LARS, Lymphoma

Acknowledgements

First and foremost, I would like to extend my sincere gratitude to my supervisor and examiner, Ida Häggström. Thank you for your invaluable guidance, insightful academic supervision, and continuous encouragement throughout this project. Your support has been instrumental in the completion of this thesis.
I am also profoundly grateful to my friends, including Zhichao Zhou and Ruiyang Han, among others. It has been a true honor to know you all. Thank you for your wonderful companionship, which makes my two years of studying abroad an unforgettable and cherished journey.
Last but certainly not least, I would like to give a special thanks to my girlfriend, Fan Lin. I am immensely lucky to have met such a lovely, kind, and gentle girl in Gothenburg. Thank you for your presence and unwavering support during the final year of my journey here. You have made this chapter of my life truly beautiful.

Yihan Tang, June 2026

Contents

List of Figures

List of Tables

1
Introduction

Cancer poses a formidable challenge to global healthcare architectures, standing as a leading cause of mortality not only in the United States [1] but also across Europe [2]. Among the myriad of oncological manifestations, lymphomas represent an exceptionally complex and heterogeneous group of malignancies. Because lymphomas [3, 4] exhibit widely diverging biological behaviors depending on their histological subtype and differentiation degree, precise staging is the absolute prerequisite for patient survival. In contemporary clinical oncology, [^{18}F]-FDG PET/CT imaging [5] has established itself as the indispensable gold standard for managing this disease, enabling clinicians to map the physiological and metabolic distribution of tumors across the entire body.

Beyond qualitative detection, doctors are increasingly relying on quantitative measures extracted from these scans, such as standardized uptake value (SUV) [6], metabolic tumor volume (MTV), and total lesion glycolysis (TLG). These metrics serve as powerful prognostic biomarkers and offer superior risk stratification compared to traditional clinical scores systems. Consequently, [^{18}F]-FDG PET/CT is no longer utilized merely for initial detection, but acts as a dynamic compass to guide therapeutic interventions and actively monitor treatment responses.

Despite the undeniable clinical value of [^{18}F]-FDG PET/CT, interpreting these scans is a highly challenging task that requires specialty expertise. When reading a scan, radiologists must carefully distinguish true tumor tissues from physiological radiotracer accumulation in normal areas [7], such as the myocardium, brain, or urinary tract. More importantly, they must identify reactive hypermetabolism – visual artifacts induced by benign inflammatory conditions or recent interventions such as chemotherapy. Doing this manually for every patient is extremely time-consuming, laborious, and highly vulnerable to subjective variability between different readers. As hospital scan volumes continue to rise, artificial intelligence (AI) powered by deep neural networks [8, 9, 10] has emerged as a promising solution to handle this heavy workload. AI tools have the potential to act as a reliable decision support tool, maintaining high diagnostic accuracy while providing results much more rapidly. A successful example is the Lymphoma Artificial Reader System (LARS) [11], which automates the detection of hypermetabolic tumor sites to achieve expert-level performance. However, while models like LARS excel at analyzing visual patterns, they still face a major clinical limitation: they operate purely on image pixels and ignore the vital patient history that a human specialist would naturally consider to rule out false positives.

In recent years, the rapid development of large language models and natural language processing has offered a new approach to breaking these vision-only limitations. In

real clinical diagnosis, physicians never pay attention to images in isolation. Instead, they always combine visual findings with patient metadata, structured medical histories, and treatment records to make comprehensive assessments. In line with this clinical logic, medical artificial intelligence is undergoing a paradigm shift from unimodal systems toward multimodal vision-language models (VLMs) [12, 13, 14]. For example, medical basic models such as BiomedCLIP [15] and PMC-CLIP [16] have successfully achieved deep semantic alignment between medical images and natural language through large-scale pre-training on massive scientific literature and image-text pairs.

However, directly applying these existing medical VLMs to lymphoma PET/CT diagnosis exists significant challenges. On the one hand, prominent medical models such as Me LLaMA [17] are pure language architectures and lack the capability to directly encode complex visual features. On the other hand, other leading medical VLMs such as BioViL [18, 19] and MedCLIP [20] have been trained almost entirely on specific radiological images, primarily chest X-rays. Given that PET/CT scans reveal the complex metabolic distribution throughout the body, directly utilizing these X-ray-centric models to extract lymphoma features would lead to a severe domain shift.

Under such conditions, the extracted features are not only uninformative but also potentially clinically misleading. Furthermore, integrating structured clinical records into deep learning pipelines presents its own architectural hurdles. While pioneering frameworks have demonstrated that linearizing metadata into natural language (e.g., TAB2TEXT [21]) enhances semantic context, a profound challenge remains: large language models process numerical data as discrete text tokens, often diluting the exact mathematical magnitude of critical continuous variables (such as exact age or precise years since initial diagnosis). How to infuse specialized visual models with dense clinical narratives without sacrificing the rigid precision of raw tabular data is a frontier that remains largely unexplored.

To bridge this gap, we introduce the Multimodal Lymphoma Artificial Reader System (MM-LARS), an end-to-end framework designed to orchestrate spatial, semantic, and numerical modalities for precise lymphoma diagnosis. Unlike generalized foundation models, MM-LARS explicitly preserves the domain-specific visual prior of the original LARS component. To overcome the loss of depth inherent in 2D Maximum Intensity Projections (MIPs), the visual backbone is adapted to ingest orthogonal dual-view scans (coronal and sagittal), endowing the network with a pseudo-3D spatial awareness necessary to resolve overlapping tissue artifacts.

Building upon this robust visual foundation, we design a dynamic tabular-to-text serialization pipeline that encodes heterogeneous patient metadata and critical therapeutic histories (such as recent chemotherapy or G-CSF administration) into highly descriptive clinical prompts. Taking inspiration from the TAB2TEXT philosophy, we advance the paradigm by proposing a tri-modal Hybrid_Attention architecture. This novel mechanism utilizes the clinical text embedding as a continuous query to cross-attend to both the spatial visual features and an explicit raw tabular vector. By doing so, the network establishes a hard mathematical gate, preserving numerical urgency while filtering out metadata noise.

Evaluated on an independent test cohort derived from a massive dataset of over

17,000 scans, MM-LARS conclusively outperforms both pure visual baselines and generalized medical VLMs. Crucially, our quantitative results demonstrate that integrating clinical context via hybrid attention optimizes the balance between sensitivity and specificity. Detailed qualitative analysis of failure cases further reveals that the network successfully mimics human clinical reasoning—mathematically "explaining away" false-positive inflammatory flares caused by recent treatments, while rescuing false-negative diagnoses in visually ambiguous presentations.
The rest of this thesis is structured as follows: Section 2 reviews related work in deep learning for PET/CT imaging, medical VLMs, and multimodal fusion strategies. Section 3 details the methodology of the MM-LARS framework, including the serialization pipeline and the evolution of our cross-attention architectures. Section 4 presents the comprehensive quantitative evaluations, ablation studies, and qualitative interpretability analyses. Finally, Section 5 discusses current limitations, outlines future research directions, and concludes the study.

2

Related Work

2.1 Deep Learning in Lymphoma Diagnosis

In recent years, deep learning has shown remarkable success in automating the analysis of [^{18}F]-FDG PET/CT scans for lymphoma patients. Early studies focused mainly on automated tumor segmentation [22, 23, 24]. These early works employed diverse architectures such as 3D U-Net [25] and its variants to automatically delineate lesions and extract quantitative metrics, such as total metabolic tumor volume (TMTV) [9, 26]. These metrics have been proved to be highly correlated with patient prognosis and survival outcomes [27].

Beyond mere segmentation, the research trajectory rapidly evolved toward end-to-end localization and disease staging. Sibille et al. [28] demonstrated the feasibility of detecting and classifying hypermetabolic uptake across whole-body scans by fusing anatomical CT and functional PET data arrays. Furthermore, Yoo [29] reviewed the Lugano classification framework, emphasizing that FDG-PET/CT has become the standard clinical tool for lymphoma staging and treatment response assessment. The author also noted that applying artificial intelligence to these imaging techniques can further improve quantitative tumor evaluation in routine clinical practice.

Building on this concept, Häggström et al. [11] developed an ensemble of ResNet34 [30] encoders on an unprecedented cohort of over 17,000 scans to automatically detect hypermetabolic tumor sites in lymphoma patients. The algorithm maintained high accuracy across both internal and external testing datasets, indicating it could serve as an effective second reader to support imaging specialists. More recently, Li et al. [31] developed a deep learning model using multi-view 2D maximum intensity projection (MIP) images from PET/CT scans to automatically stage lymphoma. This method demonstrated that orthogonal planar views can capture sufficient pseudo-3D spatial distribution while reducing processing time by several orders of magnitude.

While these image-centric models have reduced the manual workload of human radiologists significantly and improved reproducibility to a large extent, they work strictly on visual features. As a result, they remain highly susceptible to false-positive predictions. In PET/CT imaging, physiological radiotracer uptake in healthy organs (such as the myocardium, brain, or brown adipose tissue) and hypermetabolism caused by benign inflammatory conditions often mimic malignant lesions visually. In order to correctly distinguish these similarities, Human specialists typically refer to the patient's clinical background to resolve these visual ambiguities.

Due to the neglect of this essential medical history, unimodal visual networks lack

the crucial semantic context required to rule out these ambiguous artifacts, resulting in a significant gap in clinical reliability and specificity. In this work, we extend the vision-only ResNet34-based image encoder in LARS by integrating vision–language modeling to combine imaging data (PET/CT) with textual representations of clinical metadata.

2.2 Medical Vision-Language Models

In order to bridge the semantic gap between medical images and clinical concepts, VLMs have recently gained widespread attention in medical AI. Inspired by the breakthrough of contrastive language-image pretraining (CLIP) [12] in the general domain, researchers have developed various VLMs tailored for healthcare. Models such as BiomedCLIP [15] and PMC-CLIP [16] achieve broad semantic coverage by executing contrastive learning across millions of biomedical research articles, figures, and image-text pairs. Besides, architectures like BioViL [18] and MedCLIP [20] have pushed the state-of-the-art in radiological analysis by leveraging vast repositories of chest X-rays paired with expert radiology reports.

However, the direct zero-shot transfer or fine-tuning of these existing medical VLMs for functional lymphoma imaging presents insurmountable domain-shift barriers. First, models like BioViL and MedCLIP are heavily biased toward specific radiological modalities (primarily 2D anatomical radiography), which captures anatomical structures. In contrast, whole-body PET/CT scans map highly complex functional and metabolic processes throughout the whole body. Applying a chest X-ray-centric model to PET/CT would lead to a profound domain shift, which makes the extracted features uninformative. Second, large language models (LLMs) fine-tuned specifically for healthcare, such as Me LLaMA [17], focus entirely on textual understanding and natural language generation and lack the native architecture ability to directly encode raw medical image pixels.

Therefore, a domain-specific visual foundation seamlessly integrated with language modeling is still an urgent need for comprehensive PET/CT analysis. In this work, we conduct a detailed analysis and select suitable visual-language models for the diagnosis of lymphoma, and perform both qualitative analysis and quantitative comparison. By selecting the pre-trained LARS-ResNet backbone as the optimal domain-specific visual engine, we establish a robust baseline and systematically evaluate its integration with advanced textual and tabular attention mechanisms, ultimately engineering a lymphoma-specific vision-language architecture that surpasses the capabilities of generic VLMs.

2.3 Multimodal Fusion and Clinical Metadata Integration

Clinical diagnostics naturally operate as a multimodal synthesis. Expert oncologists do not interpret functional imaging in isolation; rather, they continuously cross-reference spatial visual findings with structured patient metadata (e.g., age, histo-

logical subtype, treatment phase) and unstructured clinical narratives. Within computational oncology, mapping this heterogeneous synthesis into deep learning architectures remains a persistent challenge. Conventional paradigms predominantly rely on late fusion (decision-level fusion) mechanics. A standard methodology involves extracting deep visual embeddings from convolutional networks and concatenating them with raw tabular vectors to train downstream machine learning classifiers, such as Random Forests or XGBoost [32, 33, 34]. Alternative early fusion approaches attempt to concatenate numerical metadata directly into the image feature space prior to multi-layer perceptron (MLP) processing. Yet, these naive concatenation strategies frequently fall victim to modality suppression—a phenomenon where dense, high-dimensional visual tensors overwhelm sparse, low-dimensional tabular data, thereby failing to capture true cross-modal synergy [35].

Recently, prompt-based learning has emerged as a powerful alternative. By transforming structured tabular data into descriptive natural language prompts (tabular-to-text transformation), models can leverage the rich semantic representations of pre-trained text encoders to guide the visual network via cross-attention [36]. While this linguistic transformation is highly effective for semantic alignment, emerging research suggests that discarding the raw tabular data might result in the loss of explicit numerical precision. Consequently, some cutting-edge studies have begun exploring tri-modal or hybrid architectures that simultaneously process images, text, and raw tabular features [21]. Integrating these three distinct streams requires sophisticated architectural designs, such as unified attention mechanisms, to ensure that the explicit numerical constraints of tabular data complement, rather than conflict with, the rich semantics of text and the spatial patterns of images.

Building upon these insights, our proposed MM-LARS framework systematically investigates these fusion paradigms. We explore not only prompt-driven cross-attention strategies but also hybrid fusion architectures that jointly model image, text, and tabular features. By doing so, we aim to identify the most effective mechanism to overcome modality suppression, suppress visual noise, and enhance medical image understanding in PET/CT analysis.

3

Methods

3.1 Study Design and Patients

In this study, we use a large-scale multimodal dataset comprising [^{18}F]-FDG PET/CT scans from lymphoma patients. The imaging data and corresponding clinical records are collected from the Memorial Sloan Kettering Cancer Center (MSK), New York, NY, USA. While the original LARS framework is validated on a dual-center cohort, our our research exclusively employ the MSK dataset. We make this decision because the external dataset (MUV dataset) lacks the comprehensive, structured patient metadata required to construct our multimodal tabular-to-text pipeline. The study is approved by the Institutional Review Board of MSK, and the requirement for informed patient consent is waived.

Eligible patients are those with biopsy-proven [^{18}F]FDG-avid lymphomas according to the Lugano classification guidelines [37, 38]. Specifically, the database is searched for patients with the four most common subtypes: Hodgkin lymphoma, diffuse large B-cell lymphoma, follicular lymphoma, and mantle cell lymphoma. These patients have undergone whole-body PET/CT for routine clinical purposes, such as initial staging and treatment response assessment, between January 1, 2010, and January 31, 2021. We apply strict exclusion criteria to guarantee data quality and clinical consistency. Patients are excluded if they have concurrent non-lymphoma cancers, central nervous system lymphoma, or non-[^{18}F]FDG-avid lymphomas. Furthermore, scans are excluded if the patient's blood glucose concentration exceed 200 mg/dL at the time of imaging, if non-standard PET/CT protocols are used, or if the images contain major artifacts. Consistent with standard deep learning practices for longitudinal imaging, pre-therapeutic and post-therapeutic scans from the same patient are treated as independent cases for classification.

For each included PET/CT scan, dual-view 2D MIPs, specifically coronal and sagittal views, are generated to facilitate a rapid whole-body metabolic assessment. Crucially, to construct our multimodal architecture, each imaging study is rigorously paired with its corresponding clinical metadata. This tabular data comprises 6 essential clinical features, including demographic information (e.g., age, gender) and relevant clinical history (e.g., prior chemotherapy or radiotherapy status). To rigorously prevent data leakage, the final dataset is partitioned into a training set (10606 scans), a validation set (2652 scans), and an independent test set (3325 scans) strictly at the patient level, ensuring that all scans from a single patient are confined to the same data split.

3.2 Reference Standard and Data Preprocessing

Establishing a robust ground truth and ensuring standardized data inputs are critical for training our multimodal network. In clinical practice, the metabolic response of lymphoma is internationally standardized using the 5-point Deauville score (DS), which assesses tumor [^{18}F]-FDG uptake relative to physiological background uptakes. The criteria range from DS 1 (no uptake above background) to DS 5 (uptake markedly higher than the liver and/or new lesions). For our classification task, we strictly follow standard clinical guidelines: scans scored DS 1 to 3 are classified as negative (absence of active disease), while scans scored DS 4 or 5 are classified as positive. All ground truth labels are extracted from official radiological reports reviewed by board-certified nuclear medicine physicians.
We apply a customized preprocessing pipeline to both the imaging and tabular streams to prepare the data for our MM-LARS framework.

Image Preprocessing: The raw 2D coronal and sagittal maximum intensity projections (MIPs) contain extensive empty background spaces. To optimize the field of view, we first apply a dynamic bounding box extraction algorithm to crop out the zero-value background rows and columns. The cropped images are then symmetrically padded to form a square to prevent anatomical distortion, and subsequently resized to a standardized resolution of 224×224 pixels using area interpolation. To harmonize the varying SUVs across different scans, pixel intensities are clipped to a maximum SUV of 30.0, followed by z-score normalization using the dataset's pre-calculated global mean ($\mu = 2.13$) and standard deviation ($\sigma = 3.39$). Finally, the single-channel PET projections are duplicated across three channels to ensure compatibility with pre-trained vision encoders. During the training phase, dynamic data augmentation is applied with an 80% probability to enhance generalization. Our data augmentation includes random horizontal flipping, subtle rotations (between $-10°$ and $+10°$), and the addition of Gaussian noise.

Clinical Metadata Processing: To maximize the utility of patient history, we divide the clinical metadata pipeline into two distinct processing pathways: semantic serialization and numerical encoding.

- **Semantic Pathway (Prompt Generation):** To generate the natural language prompts, we use a comprehensive merged clinical dataset. This rich repository contain over ten patient attributes, including demographic details (age, gender), disease history (days since diagnosis, smoking history, ICD-O histology descriptions, and clinical indications for the study), as well as 30-day treatment histories (chemotherapy, radiotherapy, surgery, and G-CSF administration). These diverse raw clinical descriptors are directly serialized into highly descriptive natural language prompts, preserving their full clinical nuance and unstructured semantic context for the text encoder.
- **Numerical Pathway (Tabular Feature Extraction):** Conversely, for direct ingestion into the tri-modal hybrid fusion modules, a focused subset of 6 essential clinical features is chose and subjected to rigorous quantitative preprocessing. Categorical variables are mapped to binary integers. Missing

clinical values are consistently imputed with zeros to signify the absence of an event or record (e.g., indicating no prior chemotherapy). To ensure gradient stability during deep learning optimization, these numerical variables are standardized using z-score normalization based exclusively on the mean and standard deviation of the training distribution, thereby strictly preventing any information leakage from the validation or test sets.

3.3 The MM-LARS Architecture

To effectively synthesize highly specialized metabolic visual features with heterogeneous clinical patient context, we propose the Multimodal Lymphoma Artificial Reader System (MM-LARS). The comprehensive macro-architecture of the proposed framework is illustrated in Figure 3.1. The pipeline follows a strict left-to-right data flow, structured into four sequential phases: multimodal data input, feature encoding with cross-modal dimensional alignment, text-guided tri-modal active fusion, and deep regularized classification.

3.3.1 Clinical Tabular-to-Text Serialization Pipeline

To make full use of the powerful semantic alignment ability of pre-trained language models, we first convert structured patient metadata into unstructured natural language via a dynamic tabular-to-text serialization pipeline. Rather than simply concatenating discrete text labels, we carefully design a comprehensive clinical prompt template that simulates the high-density information handover typical of radiology workflows.

According to the semantic pathway described in Section 3.2, our optimal prompt template incorporates five relevant clinical dimensions: demographic information, smoking history, histological subtype, the clinical indication for the scan, and recent therapies. The finalized prompt structure is defined as follows:

"A [age]-year-old [sex] patient with a [smoking_status] smoking history. The primary diagnosis is [histology]. The patient is undergoing PET/CT scan for [reason_for_study]. Time since initial diagnosis: [time]. The patient underwent [treatment_list] within 30 days prior to the scan."

To illustrate, a fully instantiated prompt generated for a real patient in our cohort is as follows:

"A 62-year-old male patient with a previous use smoking history. The primary diagnosis is mal lymphoma, large cell, diffuse. The patient is undergoing PET/CT scan for follow-up. Time since initial diagnosis: 1.2 years. The patient underwent chemotherapy within 30 days prior to the scan."

A major challenge in processing real-world clinical datasets is the inevitable presence of missing variables (NaNs). To maintain linguistic fluency and prevent the generation of nonsensical prompts, our pipeline incorporates a robust, rule-based fallback mechanism during the tabular-to-text transformation:

- **Missing Smoking History:** If the `Smoking_Status` attribute is missing, the pipeline dynamically omits the smoking history sentence entirely.

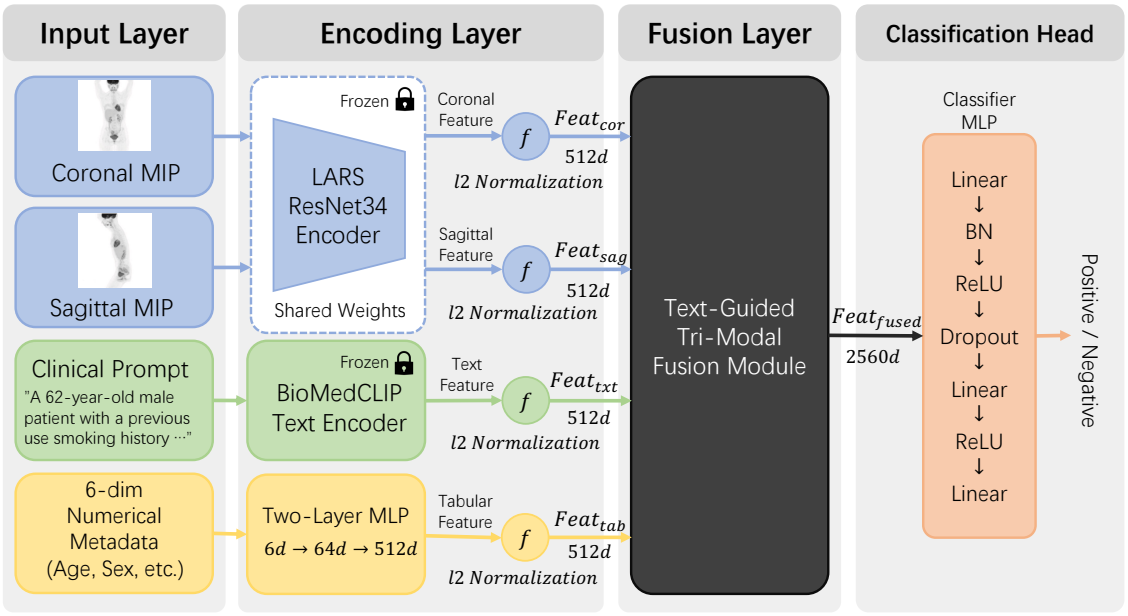


Figure 3.1: Overall macro-architecture of the Multimodal Lymphoma Artificial Reader System(MM-LARS). Heterogeneous patient data streams—dual-view PET maximum intensity projections (MIPs), clinical textual prompts, and structured numerical metadata—are processed by modality-specific encoders. The visual representations are extracted via a shared-weight, pre-trained LARS ResNet34 encoder. Unstructured clinical semantics are encoded using a frozen BioMedCLIP text encoder, while continuous and categorical tabular data are projected into a high-dimensional space via a two-layer multi-layer perceptron (MLP). Following L_2 normalization, the aligned 512-dimensional embeddings are ingested by the Text-Guided Tri-Modal Fusion Module (detailed in Figure 3.2) to construct a comprehensive 2560-dimensional joint representation. This fused feature vector is subsequently evaluated by a robust classifier MLP to predict active lymphoma involvement.

- **Missing Clinical Indication:** If the `Reason_for_Study` is unavailable, the pipeline falls back to a generalized clinical statement, substituting the specific sentence with *"The patient is undergoing PET/CT evaluation for lymphoma"*.
- **Absence of Recent Treatment:** In cases where the patient has no active therapies within the past 30 days, the treatment sentence is elegantly reformulated to state that the patient *"received no recent chemotherapy, radiation, surgery, or G-CSF medication within 30 days prior to the scan"*.

This dynamic transformation process serializes disparate, incomplete clinical data points into a cohesive, uninterrupted narrative. As shown in the middle branch of Figure 3.1, these meticulously constructed textual prompts are tokenized and processed by a frozen, pre-trained BioMedCLIP text encoder, yielding a high-dimensional clinical semantic embedding designated as $Feat_{txt} \in \mathbb{R}^{512}$.

3.3.2 Multi-Modal Feature Encoding and Dimensional Alignment

While generalized VLMs offer powerful image-text alignment capabilities, their visual encoders are predominantly pre-trained on natural images or 2D X-rays. Applying them directly to whole-body functional imaging risks severe domain shift. To preserve the highest fidelity of metabolic feature extraction, MM-LARS abandons generic visual backbones and instead utilizes a domain-specific visual foundation inherited from the original LARS architecture.

As illustrated in the upper branch of Figure 3.1, the visual backbone is built upon a customized ResNet34 architecture pre-trained on a massive cohort of lymphoma [^{18}F]-FDG PET/CT scans. The initial convolutional layer (`conv1`) is redesigned to directly ingest single-channel PET MIPs, eliminating the redundant need for three-channel duplication. By removing the last fully-connected classification head, the network outputs a 512-dimensional latent feature embedding with high density.

To capture the holistic 3D metabolic distribution without excessive computational costs, MM-LARS processes the dual-view 2D MIPs using a sequential, weight-sharing scheme. Both the coronal and sagittal projections are passed sequentially through the exact same ResNet34 network, ensuring that the spatial and metabolic features from both anatomical views are projected into a strictly aligned latent space, yielding $Feat_{cor} \in \mathbb{R}^{512}$ and $Feat_{sag} \in \mathbb{R}^{512}$.

Concurrently, to process the structured numerical clinical profiles (shown in the bottom branch of Figure 3.1), a subset of 6 essential clinical features is isolated and routed through a Two-Layer Multi-Layer Perceptron (MLP). This module gradually upscale the sparse numerical data from its native 6-dimensional space to a intermediate 64-dimensional latent space, and ultimately to a 512-dimensional aligned space, generating $Feat_{tab} \in \mathbb{R}^{512}$.

Crucially, to establish numerical parity and guarantee gradient stability prior to cross-modal interaction, all four extracted feature streams ($Feat_{cor}$, $Feat_{sag}$, $Feat_{txt}$, and $Feat_{tab}$) independently undergo L_2 normalization, which is explicitly represented by the gray circular nodes marked with "L_2" in Figure 3.1.

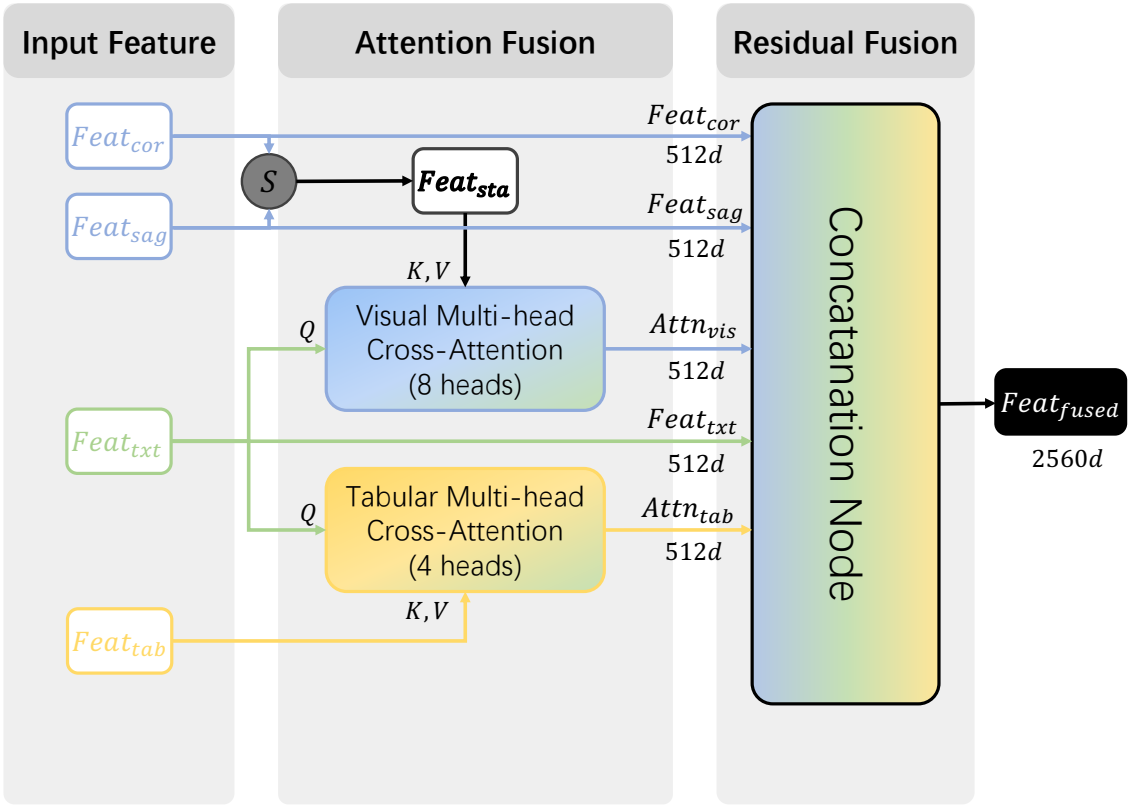


Figure 3.2: Detailed micro-architecture of the Text-Guided Tri-Modal Fusion Module. The module implements a dual-pathway cross-attention mechanism where the explicit clinical text embedding ($Feat_{txt}$) acts as the central cognitive query (Q). In the visual pathway, the text queries the stacked coronal and sagittal spatial features (K, V) to dynamically re-weight visual attention. Concurrently, in the tabular pathway, the text queries the projected numerical metadata (K, V) to filter out irrelevant clinical noise. To prevent gradient vanishing and preserve foundational modality signals, the two attention-filtered outputs ($Attn_{vis}$ and $Attn_{tab}$) are residually concatenated with the original base embeddings ($Feat_{cor}$, $Feat_{sag}$, and $Feat_{txt}$), culminating in the final 2560-dimensional fused representation ($Feat_{fused}$).

3.3.3 Text-Guided Tri-Modal Fusion Module

The core innovation of MM-LARS lies within its centralized fusion hub, represented as the shaded block in Figure 3.1. To thoroughly eliminate feature noise from the tabular stream and mitigate modality suppression, we introduce a dual-pathway, text-guided tri-modal attention network (`LARS_ResNet_Hybrid_Attention`). The deep micro-architecture and inner tensor-routing mechanics of this module are magnified and detailed in Figure 3.2.

Rather than passively concatenating features, the module establishes the clinical text embedding ($Feat_{txt}$) as the central cognitive processing unit for the model. As detailed in Figure 3.2, $Feat_{txt}$ is split into two separate pathways, which act as the global *Query* (Q) tensor to actively querying the other modalities:

- **Visual Attention Pathway:** The text query (Q) targets the visual domain to dynamically guide spatial focus. The L_2-normalized visual embeddings ($Feat_{cor}$ and $Feat_{sag}$) are first stacked along the sequence dimension via a dedicated stacking operator (denoted as node $\mathcal{S}$ in Figure 3.2) to form a unified spatio-metabolic tensor $Feat_{sta} \in \mathbb{R}^{2\times 512}$. In this pathway, $Feat_{txt}$ serves as the Query (Q), while $Feat_{sta}$ serves as both the Key (K) and Value (V). They are projected into distinct subspaces using learnable weight matrices W_Q^{vis}, W_K^{vis}, and W_V^{vis}. The scaled dot-product attention for each head is computed as:

$$\text{Attention}_{vis}(Q,K,V) = \text{softmax}\left(\frac{(Feat_{txt}W_Q^{vis})(Feat_{sta}W_K^{vis})^T}{\sqrt{d_k}}\right)(Feat_{sta}W_V^{vis})$$

  where d_k is the scaling factor representing the hidden dimension of the key. By employing a Multi-Head Cross-Attention (MHCA) mechanism with 8 independent heads, the network allows different representation subspaces to jointly attend to various clinical narratives (e.g., recent chemotherapy), outputting a dynamically filtered visual attention tensor $Attn_{vis} \in \mathbb{R}^{512}$:

$$Attn_{vis} = \text{MHCA}_{\text{8-heads}}(Feat_{txt}, Feat_{sta}, Feat_{sta})$$

- **Tabular Attention Pathway:** Concurrently, the same text query (Q) targets the tabular domain to automatically gate irrelevant numerical variables. Here, $Feat_{txt}$ continues to act as the Query (Q), while the 512-dimensional aligned tabular representation ($Feat_{tab}$) serves as the Key (K) and Value (V). Following a similar projection paradigm with specific tabular weight matrices (W_Q^{tab}, W_K^{tab}, W_V^{tab}), the tabular attention is formulated as:

$$\text{Attention}_{tab}(Q,K,V) = \text{softmax}\left(\frac{(Feat_{txt}W_Q^{tab})(Feat_{tab}W_K^{tab})^T}{\sqrt{d_k}}\right)(Feat_{tab}W_V^{tab})$$

  This pathway utilizes a separate Multi-Head Cross-Attention mechanism driven by 4 attention heads. This contextual gating neutralizes clinical noise (e.g., down-weighting the age variable if the specific histology context renders it insignificant), producing the refined tabular attention tensor $Attn_{tab} \in \mathbb{R}^{512}$:

$$Attn_{tab} = \text{MHCA}_{\text{4-heads}}(Feat_{txt}, Feat_{tab}, Feat_{tab})$$

Global Residual Concatenation: To prevent gradient vanishing during backpropagation and preserve the fundamental single-modality signals, the module enforces a strict global residual concatenation strategy. As explicitly diagrammed in Figure 3.2, a total of five distinct feature streams are horizontally channeled into the final Concatenation Node: the original base embeddings ($Feat_{cor}$, $Feat_{sag}$, and $Feat_{txt}$) and the two dynamic attention outputs ($Attn_{vis}$ and $Attn_{tab}$). Crucially, the raw tabular feature ($Feat_{tab}$) is bypassed and excluded from this final concatenation layer to prevent tabular noise injection. This fusion process culminates in an information-rich, 2560-dimensional joint representation ($Feat_{fused} \in \mathbb{R}^{2560}$), satisfying:

$$\text{Feat}_{\text{fused}} = \text{Concat}(\text{Feat}_{\text{cor}}, \text{Feat}_{\text{sag}}, \text{Feat}_{\text{txt}}, \text{Attn}_{\text{vis}}, \text{Attn}_{\text{tab}})$$

3.3.4 Deep Regularized Classification Head

The resulting 2560-dimensional fused vector is subsequently ejected from the fusion module and routed into the final classification head, as illustrated on the far-right section of Figure 3.1. To stably map this high-dimensional multimodal space to a binary diagnostic probability while strictly preventing overfitting, we implement a deep, highly regularized multi-layer perceptron (Classifier MLP).

The representation is passed through a tightly controlled sequence of operations structured as follows: a primary dense layer projects the 2560-dimensional input down to 512 dimensions, followed by a one-dimensional Batch Normalization layer to stabilize internal covariate shift, and a Rectified Linear Unit (ReLU) activation function. To enhance network robustness and enforce feature sparsity, a dropout layer with a probability of 0.3 is applied. The features are then funneled through a secondary dense layer, a second ReLU activation, and a final single-neuron projection layer to output the logit score, which is activated via a sigmoid function to yield the final binary prediction for active lymphoma involvement ("Positive" vs. "Negative").

3.4 Systematic Evolution and Variants of Fusion Mechanisms

To investigate the interaction dynamics between heterogeneous modalities and to validate the architectural superiority of the proposed MM-LARS framework, this study follows a progressive scientific exploration strategy. We design and mathematically formulate four representative fusion variants. These variants traverse from simple linear concatenations to complex attention mechanisms, serving as the theoretical foundation for our comprehensive ablation studies presented in Section 4.

3.4.1 Early Fusion Variant via Feature Concatenation

Following Occam's razor, we adopt the most straightforward deep fusion approach as our primary baseline (`LARS_ResNet_BioMedText_Base`). In this variant, cross-modal attention is entirely omitted. The L_2-normalized visual representations ($Feat_{cor}$, $Feat_{sag}$) and the text embedding ($Feat_{txt}$) are flattened and directly concatenated along the channel dimension. This early-fusion approach effectively limits the trainable parameter space and is formulated as:

$$\text{Feat}_{\text{base}} = \text{Concat}(\text{Feat}_{\text{cor}}, \text{Feat}_{\text{sag}}, \text{Feat}_{\text{txt}}) \in \mathbb{R}^{1536}$$

The 1536-dimensional vector is subsequently projected through a standard MLP classifier.

3.4.2 Text-Guided Single-Pathway View Attention Variant

To evaluate the specific contribution of text-to-vision interactions, we engineer a single-pathway attention variant (`LARS_ResNet_BioMedText_Attention`). This architecture excludes the tabular numerical data entirely. It implements only the Visual Attention Pathway described in Section 3.3.3, utilizing $Feat_{txt}$ to query the stacked spatial features $Feat_{sta}$. To maintain residual modality signals, the attention output is concatenated with the base textual and visual features:

$$\text{Feat}_{\text{view}} = \text{Concat}(\text{Feat}_{\text{cor}}, \text{Feat}_{\text{sag}}, \text{Feat}_{\text{txt}}, \text{Attn}_{\text{vis}}) \in \mathbb{R}^{2048}$$

3.4.3 Tri-Modal Hybrid Fusion Variant without Attention

To assess whether explicit numerical precision can enhance pure unstructured semantics without complex attention gating, we design a hybrid early-fusion variant (`LARS_ResNet_Hybrid_Fusion`). In this architecture, the 6-dimensional tabular vector is projected into a compact 64-dimensional latent space via a single-layer BatchNorm-activated MLP ($Feat_{tab\_64} \in \mathbb{R}^{64}$). This structured numerical representation is then aggressively concatenated with the macroscopic 512-dimensional embeddings:

$$\text{Feat}_{\text{hybrid}} = \text{Concat}(\text{Feat}_{\text{cor}}, \text{Feat}_{\text{sag}}, \text{Feat}_{\text{txt}}, \text{Feat}_{tab\_64}) \in \mathbb{R}^{1600}$$

This yields a 1600-dimensional joint vector, allowing the network to simultaneously evaluate explicit numerical boundaries and implicit text semantics, but may lead to potential feature noise due to the lack of attention filtering.

3.4.4 Decision-Level Late Fusion Variant

Finally, to provide a comprehensive methodological benchmark against traditional practices, we evaluate a decision-level late fusion strategy (`LARS + XGBoost Late Fusion`). Unlike our end-to-end differentiable networks, this paradigm decouples feature extraction from classification. The 1024-dimensional deep visual features ($Feat_{cor}$ and $Feat_{sag}$) extracted by the frozen LARS encoder are concatenated with

the raw tabular metadata post-extraction. An extreme gradient boosting (XGBoost) classifier is then trained on this combined feature set. This variant establishes a critical baseline for comparing deep multimodal synergy against conventional machine learning aggregation.

4

Experiments and Results

4.1 Experimental Settings and Evaluation Metrics

4.1.1 Implementation Details

All deep learning experiments, including model training, validation, and inference, are conducted on a remote server equipped with two NVIDIA RTX A6000 graphics processing units (48 GB VRAM). This hardware configuration is selected to accommodate the high-dimensional multimodal tensor operations and cross-attention mechanisms required by the proposed architecture. The algorithms and neural network pipelines are implemented using the PyTorch deep learning framework.

To guarantee a fair comparison across all ablation variants, we maintain strict consistency regarding the dataset partitioning. The original LARS dataset provides ten distinct cross-validation splits alongside a fixed independent test set. For our training and validation phases, we exclusively use the first data split (Fold-1). This approach not only ensures experimental reproducibility but also aligns our baseline evaluations with established literature benchmarks. Regarding the initialization of the visual feature extractor, we do not train the network from scratch. Instead, we initialize the ResNet34 visual encoder using the pre-trained weights from the top-performing individual model within the LARS-avg ensemble, as reported in the original LARS study. This domain-specific initialization strategy effectively transfers rich metabolic priors derived from a large-scale lymphoma [^{18}F]-FDG PET/CT cohort, significantly accelerating network convergence and stabilizing the subsequent multimodal fusion process.

During the training phase, the batch size is uniformly set to 64, with a maximum limit of 30 epochs. Network optimization is driven by the AdamW (Adaptive Moment Estimation with Weight Decay) optimizer. We set the initial learning rate to 1×10^{-5} and applied a weight decay coefficient of 1×10^{-4}. The choice of AdamW is motivated by its decoupled weight decay mechanism, which provides superior regularization for deep, highly parameterized networks, thereby mitigating the risk of overfitting on the relatively constrained medical dataset. Additionally, we implement a dynamic early stopping strategy to capture the optimal generalization state of the models and prevent unnecessary computational expenditure. The system continuously monitor the Area Under the Curve (AUC) on the validation set. If the validation AUC failed to improve by a minimum delta of 0.0001 for 5 consecutive epochs, the training process was automatically halted. The model checkpoint

with the highest validation AUC is preserved as the optimal configuration for final testing.

4.1.2 Evaluation Metrics and Threshold Determination Strategy

The quantitative assessment of the proposed models primarily rely on three standard clinical classification metrics: AUC, Sensitivity (also referred to as Recall or True Positive Rate), and Specificity (True Negative Rate).
A persistent challenge in evaluating predictive models on imbalanced clinical datasets is the selection of an appropriate classification threshold. The conventional practice of defaulting to a 0.5 probability cutoff often yields heavily skewed predictions, leading to either unacceptable false-negative rates (missed active lymphoma diagnoses) or inflated false-positive rates (over-diagnosis due to benign inflammation). To counteract this intrinsic bias, our evaluation protocol includes a dynamic threshold optimization strategy based on Youden's Index (J).
For each trained model variant, the predicted probabilities are first generated for the validation set. We compute the True Positive Rate (TPR) and False Positive Rate (FPR) across the entire spectrum of possible operating thresholds. Youden's Index is then calculated as $J = TPR - FPR$. The specific probability value that maximized this index is identified as the optimal diagnostic threshold. Crucially, to prevent data leakage and ensure an unbiased final evaluation, this optimal threshold is strictly locked. It is subsequently applied to the independent, unseen test set to binarize the continuous logit predictions, enabling the construction of the final confusion matrix and the precise, clinically relevant calculation of test-set Sensitivity and Specificity.
Note that the quantitative metrics reported in this study are derived from the single deterministic evaluation of the best-performing model checkpoint. Specifically, for each architectural variant, the single checkpoint that achieves the highest AUC on the validation set during training is selected and subsequently applied to the independent test set. These results do not represent an ensemble or an average across multiple stochastic runs.

4.2 Optimization of Single-Modality Foundations

Before constructing the complex multimodal cross-attention networks, it is imperative to independently evaluate and optimize the performance of individual modality foundations. All ablation experiments in this section are conducted on the validation set. The primary objective is to establish the optimal spatial visual input strategy and identify the most robust image and text feature extractors, thereby laying a solid empirical foundation for the subsequent multimodal fusion phases.

4.2.1 Single-View vs. Dual-View Feature Extraction

Lymphoma lesions exhibit highly heterogeneous and complex three-dimensional spatial distributions across the whole body. In standard clinical practice, rendering 3D volumetric PET scans into 2D MIPs inevitably leads to the loss of depth information.

Consequently, a single projection angle frequently suffers from severe anatomical overlap, where malignant lesions may visually merge with adjacent normal organs that exhibit high physiological [^{18}F]-FDG uptake (such as the myocardium, brain tissue, or urinary tract).
To quantitatively investigate the impact of different projection angles on diagnostic performance, we conduct a spatial ablation study using the isolated visual baseline model (`LARS_ResNet_VisionOnly`). The experiment evaluate three distinct input configurations: Coronal-Only, Sagittal-Only, and the integrated Dual-View. The corresponding evaluation metrics on the validation set are summarized in Table 4.1.

Table 4.1: Performance comparison of different spatial view inputs on the validation set using the `LARS_ResNet_VisionOnly` baseline.

Spatial Input Strategy	Feature Dimension	Validation AUC	Sensitivity	Specificity
Coronal-Only	512	0.9216	0.7888	0.8911
Sagittal-Only	512	0.9167	0.7976	0.8840
Dual-View (Coronal + Sagittal)	1024	**0.9407**	0.7642	0.9414

Table 4.1 reveals a distinct performance bottleneck when the network relies solely on a single projection. The Coronal-Only model achieves a respectable AUC of 0.9216, offering a broad lateral field of view; however, it is constrained by a relatively low specificity of 0.8911 due to anterior-posterior tissue overlap. Conversely, the Sagittal-Only model provides anterior-posterior depth but loses bilateral symmetry, resulting in a slightly lower AUC of 0.9167 and the lowest specificity (0.8840) among the three configurations.
The synergistic integration of both orthogonal projections in the Dual-View architecture yields a substantial performance leap, achieving the highest overall discriminative power with an AUC of 0.9407. Most notably, the Dual-View approach drives a remarkable surge in Specificity, peaking at 0.9414. This significant improvement effectively demonstrates the clinical mechanism of the dual-pathway input: by cross-referencing orthogonal planes within the fully connected layers, the network successfully acquires a "pseudo-3D" spatial awareness. This structural comprehension enables the model to accurately disambiguate true hypermetabolic malignancies from overlapping benign physiological artifacts, drastically reducing the false-positive rate.
While a marginal decrease in Sensitivity (0.7642) is observed—a common trade-off when a model adopts a more conservative, high-specificity decision boundary—the substantial gain in the overall AUC confirms the net diagnostic superiority of the orthogonal integration. Consequently, the Dual-View configuration is definitively established as the standard visual input pipeline for all subsequent multimodal architectures in this study.

4.2.2 General Foundation Models vs. Domain-Specific Architecture

Having established the diagnostic superiority of the Dual-View input strategy, this section systematically evaluates the core "visual engine" of the framework—the

image feature extractor. In recent years, VLMs pre-trained on massive medical image-text pairs have demonstrated remarkable cross-modal generalization. However, whether these general-purpose foundation models can be directly transferred to highly specialized functional imaging modalities, such as [^{18}F]-FDG PET/CT, without suffering from severe domain shift remains an open empirical question.
To rigorously investigate this, we conduct a comparative ablation study evaluating our proposed domain-specific visual backbone (`LARS_ResNet`) against the visual encoders of two prominent open-source medical VLMs: PMC-CLIP and BioMedCLIP. To maintain strict control variables, all three feature extractors are evaluated under the optimal Dual-View input configuration without any downstream text fusion. Their quantitative performance on the validation set is detailed in Table 4.2.

Table 4.2: Performance comparison of different visual feature extractors on the validation set.

Visual Encoder Backbone	Dominant Pre-training Domain	Validation AUC
PMC-CLIP Vision Encoder	Biomedical literature figures & general medical images	0.7818
BioMedCLIP Vision Encoder	Large-scale PubMed image-text pairs (anatomical heavy)	0.8252
LARS_ResNet	**Million-scale specialized lymphoma PET/CT cohort**	**0.9495**

The results in Table 4.2 clearly illustrate that transferring general medical vision encoders directly to this specialized functional imaging task yields suboptimal diagnostic outcomes. The visual encoder of PMC-CLIP demonstrates the lowest discriminative power, yielding an AUC of only 0.7818. This underperformance can be primarily attributed to its pre-training corpus, which predominantly consists of 2D illustrations, charts, and macroscopic clinical photographs extracted from biomedical literature. Consequently, the network inherently lacks the representational depth required to process high-resolution, whole-body radiological scans.
While BioMedCLIP—pre-trained on a more extensive dataset encompassing a substantial volume of radiological imaging (such as chest X-rays)—achieves a moderately improved AUC of 0.8252, it still falls significantly short of clinical applicability thresholds. This performance gap exposes a critical limitation in cross-modal transfer: the visual latent spaces of general VLMs are predominantly anchored in recognizing "anatomical structures" (e.g., skeletal boundaries, organ contours, and tissue opacities). Conversely, the diagnostic interpretation of lymphoma PET imaging relies fundamentally on identifying "abnormal glucose metabolism"—functional anomalies indicated by radiotracer distribution, rather than mere structural deformations. General visual foundations struggle to reliably capture these subtle, highly specific metabolic variations, inevitably leading to an insurmountable performance bottleneck.
In stark contrast, the domain-specific `LARS_ResNet` exhibits an overwhelming advantage, driving the validation AUC to an impressive 0.9495. This substantial performance leap (an approximate 15% improvement over BioMedCLIP) validates the necessity of domain-specific pre-training. By undergoing deep feature regularization on a massive cohort of specialized lymphoma PET/CT scans, the `LARS_ResNet` weights possess exceptionally high fidelity in extracting disease-specific metabolic hotspots, recognizing benign physiological artifacts, and delineating tumor hetero-

geneity. Ultimately, these empirical findings highlight the severe domain shift limitations inherent in generic VLMs for complex functional imaging tasks, thereby firmly justifying the selection of the custom `LARS_ResNet` as the foundational visual backbone for the ensuing MM-LARS multimodal framework.

4.2.3 Clinical Prompt Engineering

Having determined the optimal visual foundation, the final phase of our single-modality optimization focuses on the textual input. In VLMs, the quality of the cross-modal alignment heavily depends on the semantic richness and structural integrity of the input text. To investigate how varying degrees of clinical context influence the network's diagnostic behavior, we conduct a systematic prompt engineering ablation study.

For this experiment, we adopt the `BioMedCLIP_Dual_Attention` architecture to isolate the impact of textual phrasing while maintaining a consistent visual baseline. We evaluate four distinct prompt versions (ranging from `prompt_v1` to `prompt_v4`), representing a progressive increase in clinical information density—from basic demographic templates to highly serialized, comprehensive patient histories (as previously detailed in Section 3.3.1). The quantitative results on the validation set are presented in Table 4.3.

Table 4.3: Performance evaluation of progressive clinical prompt versions on the validation set.

Prompt Version	Clinical Context Included	Validation AUC	Sensitivity	Specificity
`prompt_v1`	Demographics & Reason for study	0.8353	**0.7625**	0.7493
`prompt_v2`	v1 + Histological subtype	0.8359	0.7016	0.8216
`prompt_v3`	v2 + Smoking history	0.8329	0.6557	0.8574
`prompt_v4`	**v3 + Recent treatments & Timeline**	**0.8469**	0.6712	**0.8671**

A progressive analysis of the data in Table 4.3 reveals a fascinating shift in the model's diagnostic decision boundaries driven entirely by semantic granularity.

When utilizing `prompt_v1`, which provides only rudimentary patient metadata (e.g., age, sex), the model achieves a baseline AUC of 0.8353. Notably, this version achieves the highest sensitivity (0.7625) but a disproportionately low specificity (0.7493). Clinically, this indicates an over-reliance on ambiguous visual hotspots; without sufficient background context, the model adopts an overly aggressive classification strategy, frequently misinterpreting benign physiological uptake as active malignancy (resulting in high false-positive rates).

As we enrich the prompt with specific disease characteristics in `prompt_v2` and `prompt_v3` (incorporating histological subtypes and smoking history), the model's behavior fundamentally transforms. While overall AUC remains relatively stable, we observe a dramatic surge in specificity (climbing to 0.8574 in v3), accompanied by a natural trade-off in sensitivity. This trend convincingly demonstrates the mechanism of "contextual gating". The enriched text embeddings successfully provide a cognitive constraint, teaching the attention mechanism to suppress visual noise when the clinical narrative do not align with a high-probability lymphoma profile.

The most compelling breakthrough occurs with the implementation of `prompt_v4`. To provide an intuitive understanding of its semantic richness, an actual instantiated example from our dataset is presented as follows:

> *"A 62-year-old male patient with a previous use smoking history. The primary diagnosis is mal lymphoma, large cell, diffuse. The patient is undergoing PET/CT scan for follow-up. Time since initial diagnosis: 1.2 years. The patient underwent chemotherapy within 30 days prior to the scan."*

By explicitly integrating these crucial temporal variables and recent therapeutic interventions (such as the chemotherapy mentioned above), the overall diagnostic performance peak at an AUC of 0.8469, alongside the highest specificity of 0.8671. In the realm of ^{18}F-FDG PET/CT, recent treatments are notorious for causing reactive inflammation (such as diffuse bone marrow or splenic uptake), which acts as a major visual confounder. The explicit encoding of this treatment history in `prompt_v4` allows the text-guided attention module to mathematically "explain away" these high-uptake artifacts.

Consequently, the fully serialized `prompt_v4` not only maximizes the overall discriminative power (AUC) but also best aligns with the rigorous requirements of clinical radiology by effectively minimizing false alarms. Thus, `prompt_v4` is strictly adopted as the standard text input strategy for all subsequent multi-modal evaluations in the MM-LARS framework.

4.3 Evolution of Multimodal Fusion Strategies

Following the optimization of isolated modality inputs, this section maps the architectural trajectory of the proposed MM-LARS framework. The objective is to rigorously assess how progressively sophisticated fusion strategies—ranging from simplistic early concatenation to deep text-guided cross-attention—influence the network's learning dynamics. Table 4.4 presents the validation performance of six distinct structural variants. It is crucial to first contextualize these metrics: the validation set AUC serves primarily as an indicator of a model's upper representational bound and its ability to fit the training distribution, rather than an absolute guarantee of clinical generalizability on unseen data.

Table 4.4: Performance evolution of multimodal fusion strategies on the validation set.

Model Variant	Input Modalities	Fusion Strategy	Validation AUC
`LARS_ResNet_VisionOnly`	Vision	Visual Baseline	0.9495
`LARS_ResNet_BioMedText_Base`	Vision + Text	Early Concatenation	0.9527
`LARS_ResNet_BioMedText_Attention`	**Vision + Text**	**Cross-Attention**	**0.9535**
`LARS_ResNet_Hybrid_Fusion`	Vision + Text + Tabular	Tri-modal Concatenation	0.9473
`LARS_ResNet_Hybrid_Attention`	Vision + Text + Tabular	Tri-modal Cross-Attention	0.9471
`Late Fusion with XGBoost`	Vision + Text + Tabular	Decision-Level Ensemble	0.9397

A comparative review of the early architectural variants clearly demonstrates the diagnostic value of incorporating unstructured clinical narratives. Moving from the isolated visual baseline (AUC: 0.9495) to the early-concatenation bimodal model

(`LARS_ResNet_BioMedText_Base`, AUC: 0.9527), the network exhibits a measurable improvement, confirming that even naive vector stacking provides useful contextual anchors for the visual features. However, the true capacity of vision-language synergy emerges with the implementation of cross-attention (`LARS_ResNet_BioMedText_Attention`). By leveraging the clinical text embedding as an active cognitive query to dynamically recalibrate spatial visual distributions, this variant achieves the highest empirical performance on the validation set with an AUC of 0.9535. This suggests that the attention-driven alignment effectively pushes the network to its theoretical fitting peak for this specific dataset split.

Building upon this highly optimized bimodal latent space, the subsequent introduction of structured numerical metadata (such as age or intervals between diagnoses) introduces unexpected optimization challenges. As reflected in Table 4.4, both tri-modal architectures experience a marginal contraction in validation performance. Specifically, the `LARS_ResNet_Hybrid_Fusion` and `LARS_ResNet_Hybrid_Attention` models yield AUCs of 0.9473 and 0.9471 respectively. Rather than indicating a flaw in the network design, this slight dip highlights the inherently noisy and sparse nature of real-world tabular data. When the network is forced to directly concatenate these raw, often fluctuating numerical variables with highly refined vision-language vectors—as seen in the `Hybrid_Fusion` approach—it suffers from representational interference. The model risks memorizing metadata noise instead of extracting generalizable patterns. While the `Hybrid_Attention` architecture is explicitly engineered to mathematically gate and suppress this noise, the sheer optimization complexity of aligning three profoundly different statistical distributions still slightly constrains its absolute fitting capacity on the validation data.

To contextualize these end-to-end deep learning strategies against traditional machine learning paradigms, we also evaluate a decision-level ensemble. In the `Late Fusion with XGBoost` variant, deep visual and textual features are extracted offline and processed by a gradient boosting tree alongside the raw tabular variables. The relatively inferior performance of this disjointed approach (AUC: 0.9397) exposes a critical limitation: decoupling feature extraction from the final classification deprives the network of the ability to mutually inform and correct cross-modal errors during the backpropagation process.

Ultimately, the validation metrics establish the text-guided bimodal attention network as the architecture with the strongest raw fitting capability. However, clinical robustness demands more than just fitting the validation distribution; it requires an exceptional ability to maintain high sensitivity while rigorously suppressing false positives in ambiguous, unseen cases. Whether the sophisticated noise-gating properties of the tri-modal `Hybrid_Attention` model can translate into superior generalization and clinical reliability remains the critical question. This ultimate evaluation is the central focus of Section 4.4, where all models confront the hold-out test set.

4.4 Evaluation on the Independent Test Set

The ultimate benchmark of any medical artificial intelligence system lies not in its ability to fit a validation distribution, but in its robustness when confronted with a completely independent, unseen patient cohort. In this final quantitative

phase, we evaluate the generalization capabilities of our architectural variants on the hold-out test set. Crucially, to ensure strict clinical realism and prevent data leakage, the optimal operating thresholds for each model are determined exclusively on the validation set using Youden's Index, as detailed in Section 4.1.2. These locked thresholds are then applied to the continuous logit predictions of the test set to generate definitive confusion matrices, allowing for an unbiased assessment of Sensitivity and Specificity. The comprehensive performance metrics across all evaluated models are summarized in Table 4.5.

Table 4.5: Final quantitative evaluation on the independent hold-out test set.

Model Architecture	Test AUC	Sensitivity	Specificity
`PMC_CLIP_Dual_Attention_FT`	0.8623	0.6923	0.8638
`BioMedCLIP_Dual_Attention_FT`	0.9064	0.7796	0.9017
`Late Fusion with XGBoost`	0.9295	0.8398	0.9067
`LARS_ResNet_VisionOnly`	0.9347	0.8552	0.8865
`LARS_ResNet_BioMedText_Base`	**0.9407**	0.8687	0.8683
`LARS_ResNet_BioMedText_Attention`	0.9366	0.8345	**0.9125**
`LARS_ResNet_Hybrid_Fusion`	0.9364	**0.8894**	0.8007
`LARS_ResNet_Hybrid_Attention`	0.9403	0.8678	0.8713

An initial review of the test set metrics provides conclusive evidence regarding the necessity of domain-specific visual pre-training. Even after partial fine-tuning (FT), the architectures reliant on the visual encoders of general foundation models—namely `PMC_CLIP_Dual_Attention_FT` (AUC: 0.8623) and `BioMedCLIP_Dual_Attention_FT` (AUC: 0.9064)—continue to lag significantly behind the baseline `LARS_ResNet_VisionOnly` model (AUC: 0.9347). This persistent performance gap confirms that general-purpose VLMs still suffer from unresolved domain shifts when applied to highly specialized functional imaging, further validating our strategic choice to anchor the multimodal framework on the custom LARS backbone.

Building upon this robust visual foundation, the introduction of unstructured clinical text successfully elevates the network's discriminative power. The `LARS_ResNet_BioMedText_Base` model, which utilizes an early concatenation of visual and textual embeddings, achieves an impressive test AUC of 0.9407, demonstrating excellent balance with a sensitivity of 0.8687 and a specificity of 0.8683. Interestingly, transitioning to the cross-attention mechanism in the `BioMedText_Attention` variant leads to a more conservative diagnostic boundary; while the overall AUC slightly contracts to 0.9366, the model achieves a remarkable specificity of 0.9125. This indicates that text-guided attention inherently trains the network to be highly cautious of false positives, sacrificing a degree of sensitivity to prioritize diagnostic certainty.

However, the true complexity of multimodal clinical AI is exposed when introducing structured numerical metadata (such as age, gender, and treatment intervals) to form tri-modal architectures. When evaluating the `LARS_ResNet_Hybrid_Fusion` model—which naively concatenates the tabular vector with the vision-language embeddings—a severe degradation in clinical reliability is observed. Despite maintaining a high sensitivity (0.8894), its specificity plummets to an unacceptable

0.8007. This sharp decline perfectly corroborates the hypothesis formulated during our validation phase: forcing a network to directly ingest highly variable, raw numerical metadata without an explicit gating mechanism causes it to overfit to localized dataset noise. Consequently, when deployed on unseen patients, this memorized noise translates into a disproportionately high rate of false-positive diagnoses. In stark contrast, the fully realized `LARS_ResNet_Hybrid_Attention` architecture demonstrated exceptional clinical robustness, gracefully navigating the pitfalls of tabular noise. Guided by the explicit clinical text query, the tabular cross-attention pathway successfully filtered out numerical interference while actively retaining critical continuous variables. This sophisticated feature alignment culminated in a remarkably balanced diagnostic profile, achieving an AUC of 0.9403, alongside a Sensitivity of 0.8678 and a Specificity of 0.8713. While its AUC is statistically identical to the bimodal baseline (a negligible difference of 0.0004), the `Hybrid_Attention` model achieves a superior specificity and proves its unique structural capability to safely integrate all three clinical modalities without succumbing to the catastrophic specificity collapse seen in the naive fusion approach.
Furthermore, contextualizing these deep learning strategies against traditional machine learning paradigms further solidifies the superiority of our proposed framework. The decision-level ensemble approach, `Late Fusion with XGBoost`, achieves a respectable specificity (0.9067) but falls short in overall AUC (0.9295) and sensitivity (0.8398). This suggests that extracting deep features offline and delegating the final decision to a tree-based classifier fails to capture the intricate, non-linear synergies between modalities, reinforcing the necessity of deep, end-to-end cross-modal backpropagation.
Ultimately, the comprehensive evaluation on the independent test set definitively establishes `LARS_ResNet_Hybrid_Attention` as the optimal, most resilient multimodal architecture. By demonstrating its capacity to mathematically gate tabular noise while maintaining excellent sensitivity, the ensuing critical inquiry is to understand exactly *how* this text-guided mechanism resolves complex, ambiguous clinical scenarios at the individual patient level. To this end, Section 4.5 provides an in-depth qualitative analysis, dissecting specific diagnostic failure cases to unveil the model's unprecedented clinical interpretability.

4.5 Qualitative Analysis and Interpretability

While the quantitative evaluation in Section 4.4 conclusively demonstrates the superior discriminative power and clinical robustness of the `LARS_ResNet_Hybrid_Attention` architecture, aggregate metrics alone cannot elucidate the underlying decision-making mechanisms of the network. To bridge the gap between algorithmic performance and clinical interpretability, this section provides an in-depth qualitative analysis of specific diagnostic scenarios. By deliberately isolating cases where the baseline single-modality network fails but the multimodal architectures succeed, we can uncover exactly how clinical text and structured metadata interact with spatial visual features to correct cognitive biases.
An automated screening of the independent test set predictions reveals a compelling statistic: a total of 106 specific patient cases are incorrectly classified by the

`LARS_ResNet_VisionOnly` baseline, yet are successfully rescued and correctly diagnosed by both the bimodal (`BioMedText_Base`) and tri-modal (`Hybrid_Attention`) models. To provide rigorous context for the subsequent probability scores, it is important to recall the optimal Youden's Index decision thresholds locked during the validation phase: 0.4123 for Vision-Only, 0.3586 for the Base model, and 0.4026 for the Hybrid_Attention model.

4.5.1 Semantic Correction: Overcoming Visual Ambiguity and False Negatives

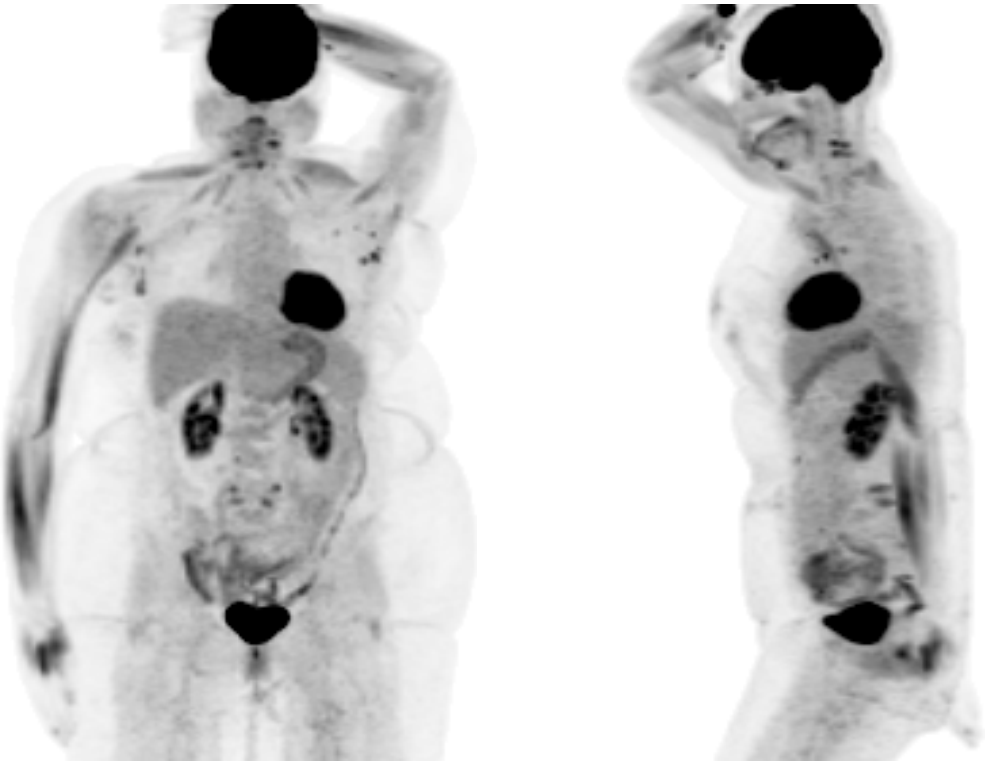
One of the most profound limitations of relying solely on MIPs in [^{18}F]-FDG PET/CT is the potential for false-negative diagnoses in the absence of focal, high-intensity radiotracer uptake. Certain histological subtypes or specific clinical phases often present with subtle, diffuse, or microscopically disseminated disease that visually blends into the background of normal physiological uptake (such as in the brain, myocardium, kidneys, and bladder). In these scenarios, semantic clinical context acts as an indispensable cognitive anchor to override visual ambiguity.

A striking pattern of this semantic rescue is evident in the diagnosis of mantle cell lymphoma, as demonstrated by the representative cases of Patient 458223 and Patient 460029. Both patients have active disease (True Label = 1). However, upon examining the corresponding coronal and sagittal MIP scans for Patient 458223 (Figure 4.1) and Patient 460029 (Figure 4.2), the visual evidence of malignancy is highly ambiguous.

As observed in these images, the radiotracer distribution appears largely physiological, lacking the prominent, hypermetabolic mass configurations typically associated with aggressive lymphomas. Consequently, the `LARS_ResNet_VisionOnly` model heavily penalizes these scans, outputting low confidence probabilities of 0.3962 (Patient 458223) and 0.2894 (Patient 460029). Because both scores falls below the locked visual threshold of 0.4123, the vision-only network incorrectly classifies them as negative.

The diagnostic trajectory completely reverses with the integration of the clinical narratives. For both patients, the serialized prompt explicitly states that the primary diagnosis is mantle cell lymphoma and that the scan is intended for initial staging. Clinically, mantle cell lymphoma frequently involves the gastrointestinal tract or bone marrow, or presents with sub-centimeter lymphadenopathy that is notoriously difficult to resolve on 2D MIPs without volumetric cross-referencing. Furthermore, the phrase "initial staging" establishes an exceptionally high pre-test probability of active disease, as the patient has just been diagnosed and requires a baseline measurement prior to any therapy. The text encoder successfully maps this rich semantic prior, altering the attention weights to look beyond the absence of massive focal tumors. Consequently, the multimodal networks drastically elevate their predictions, with the `Hybrid_Attention` model outputting highly confident, correct probabilities of 0.9485 and 0.9398, respectively.

This mechanism of text-guided semantic correction extends beyond disease subtypes to encompass the confounding effects of recent clinical interventions. Patient 458424 presents another challenging false-negative scenario (True Label = 1), where the



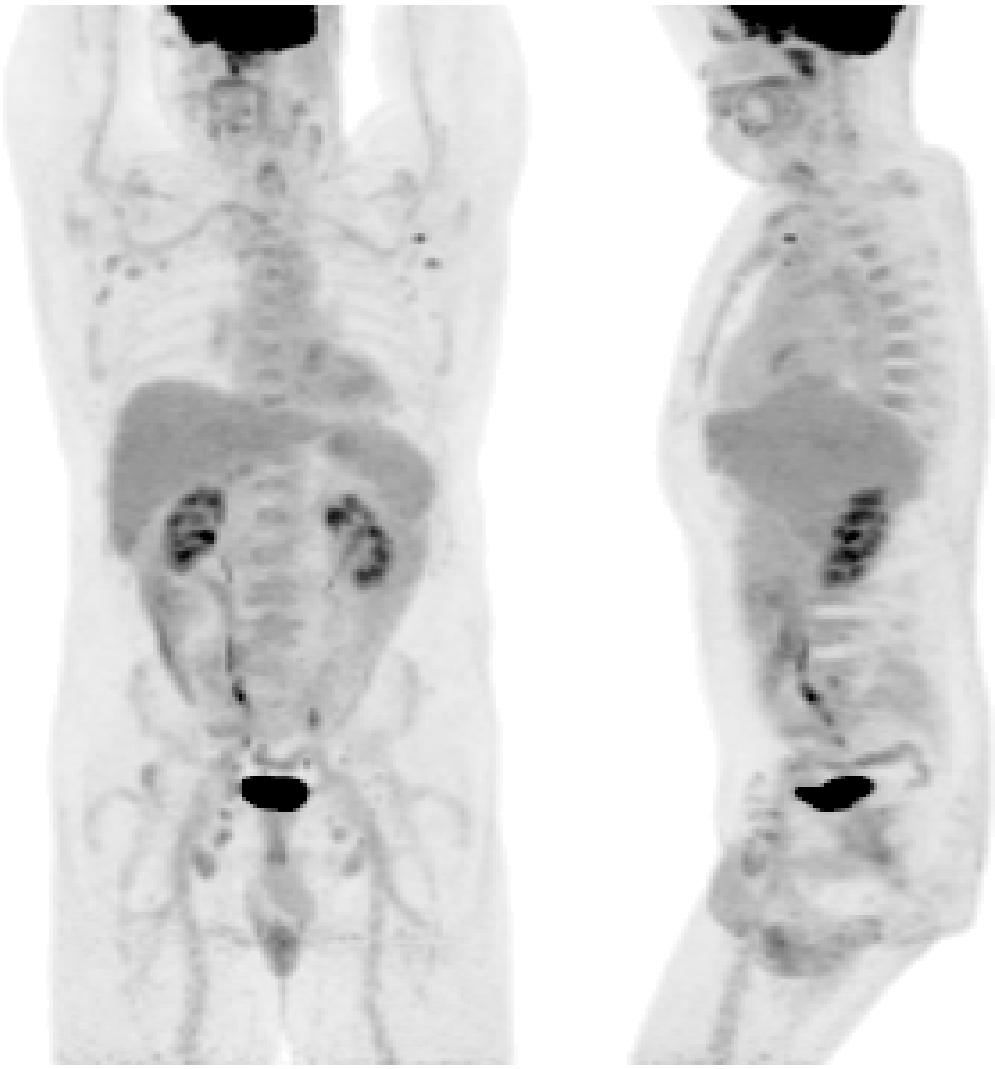
(a) Coronal view **(b)** Sagittal view

Figure 4.1: Dual-view PET maximum intensity projections (MIPs) of Patient 458223. The orthogonal coronal (left) and sagittal (right) scans display a predominantly physiological radiotracer distribution, lacking prominent hypermetabolic nodal masses, which led to a visual-only false negative.

visual features shown in **Figure 4.3** are likely confounded by systemic physiological alterations, making the active disease indistinguishable from background noise to an isolated convolutional network.
As a result, the Vision-Only model yields an insufficient probability of 0.3786, missing the diagnosis. However, the patient's textual prompt provides a critical piece of contextual evidence: *"The primary diagnosis is infiltrating duct carcinoma... The patient underwent chemotherapy within 30 days prior to the scan."* Recent chemotherapy is well-known in nuclear medicine for causing reactive marrow hyperplasia or splenic inflammation, which fundamentally alters the standard biodistribution of ^{18}F-FDG. More importantly, evaluating a scan immediately after or during recent chemotherapy implies a complex clinical state where residual active disease might be temporarily suppressed in metabolic intensity but remains uneradicated. Armed with this temporal and therapeutic context, the text-guided attention module mathematically contextualizes the visual ambiguity. By cross-attending the text embedding with the spatial features, the network successfully "explains away" the atypical visual presentation, rescuing the case with a soaring correct hybrid probability score of 0.9151.

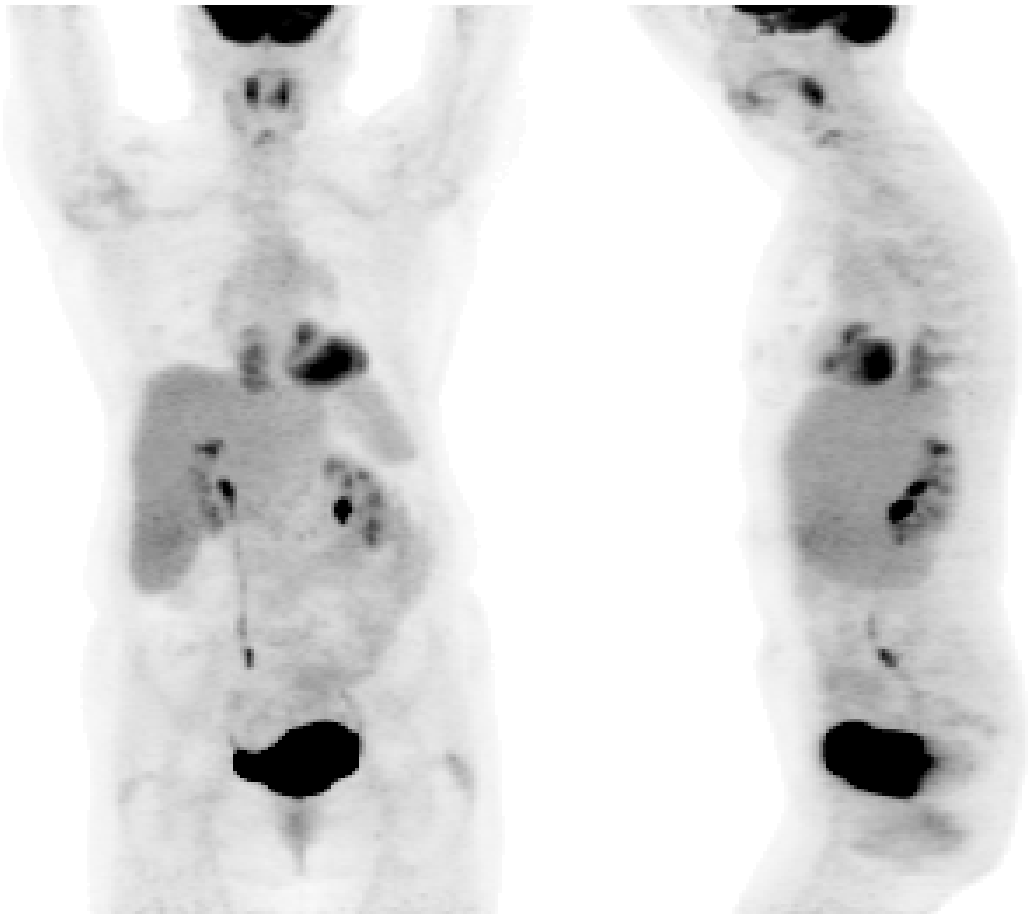
4.5.2 Tabular Attention Gating: Precision Diagnostics in Ambiguous Scenarios

While the bimodal text-guided architecture successfully mitigates numerous visual misclassifications, our evaluation reveals a subset of extremely challenging scenarios where both the isolated visual backbone and the baseline text-concatenation model (`BioMedText_Base`) fail. An automated screening of the test set identifies exactly 34 such highly evasive cases. In these instances, the visual ambiguity is so profound—and the clinical context so nuanced—that standard multimodal concatenation proved insufficient. Only the fully realized `LARS_ResNet_Hybrid_Attention` model, equipped with the tabular cross-attention pathway, manages to rescue these diagnoses successfully. To contextualize the model outputs, the previously locked optimal Youden's thresholds are applied: 0.4123 for Vision-Only, 0.3586 for the Base model, and 0.4026 for the Hybrid_Attention model.
The distinct advantage of tabular attention lies in its handling of precise numerical metadata. While language models process numbers as discrete text tokens, often diluting their continuous significance within a dense clinical narrative, the tabular pathway explicitly encodes exact numerical magnitudes (e.g., precise age or exact years since diagnosis). This allows the cross-attention mechanism to function as a strict mathematical gate.
This numerical urgency is perfectly illustrated in the false-negative rescue of Patient 392256 (True Label = 1). As depicted in Figure 4.4, the dual-view MIPs display a highly ambiguous radiotracer distribution without obvious focal lesions.
Consequently, the Vision-Only model yields a probability of 0.4005, falling short of its threshold. Surprisingly, the `BioMedText_Base` model also fails, outputting an insufficient 0.3539. Although the text prompt mentioned *"Time since initial diagnosis: 0.9 years"* and *"underwent radiation therapy within 30 days"*, the simple concatenation mechanism struggles to appropriately weight the critical numerical



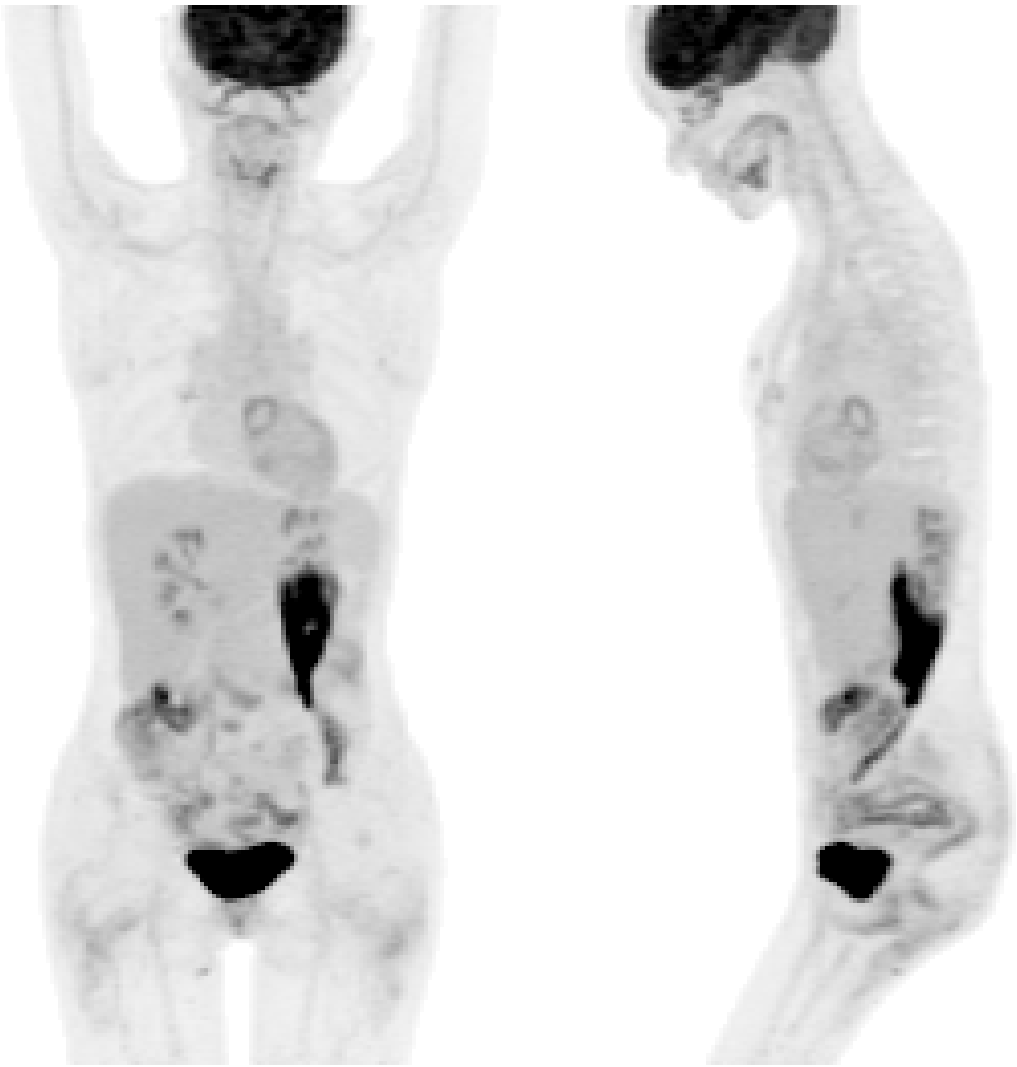
(a) Coronal view **(b)** Sagittal view

Figure 4.2: Dual-view PET maximum intensity projections (MIPs) of Patient 460029. The visual representation shows subtle, poorly defined metabolic activity that is easily confounded by normal background tracer excretion, resulting in a misclassification by the pure vision network.



(a) Coronal view **(b)** Sagittal view

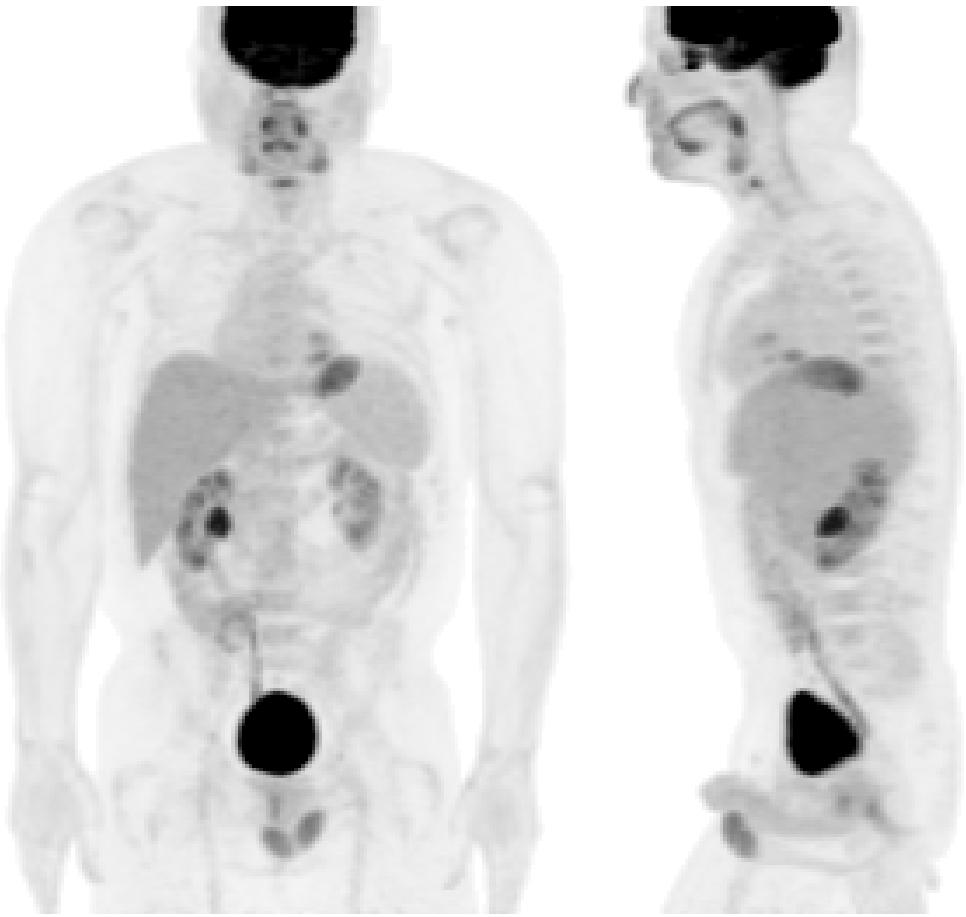
Figure 4.3: Dual-view PET maximum intensity projections (MIPs) of Patient 458424. The coronal (left) and sagittal (right) projections exhibit an altered systemic tracer biodistribution secondary to recent chemotherapy, masking the underlying active malignancy from the vision-only backbone.



(a) Coronal view **(b)** Sagittal view

Figure 4.4: Dual-view PET MIPs of Patient 392256. The coronal (left) and sagittal (right) scans exhibit diffuse, non-specific tracer uptake. The lack of prominent metabolic hotspots led both the vision-only and text-base models to generate false-negative predictions.

value of "0.9 years" against the background text. However, the `Hybrid_Attention` model directly ingests this numerical metadata as an explicit continuous feature. Recognizing that 0.9 years represents a highly vulnerable early-relapse window for aggressive diffuse large B-cell lymphoma (DLBCL), the tabular attention mechanism amplifies the localized visual signals despite the masking effect of recent radiation therapy, producing a soaring true-positive confidence score of 0.7827.
Equally critical to clinical safety is the model's ability to suppress false positives in patients with long-term remission. Patient 394559 (True Label = 0) represents a case where benign physiological or inflammatory uptake mimicking malignancy tricked the earlier networks, as shown in **Figure 4.5**.

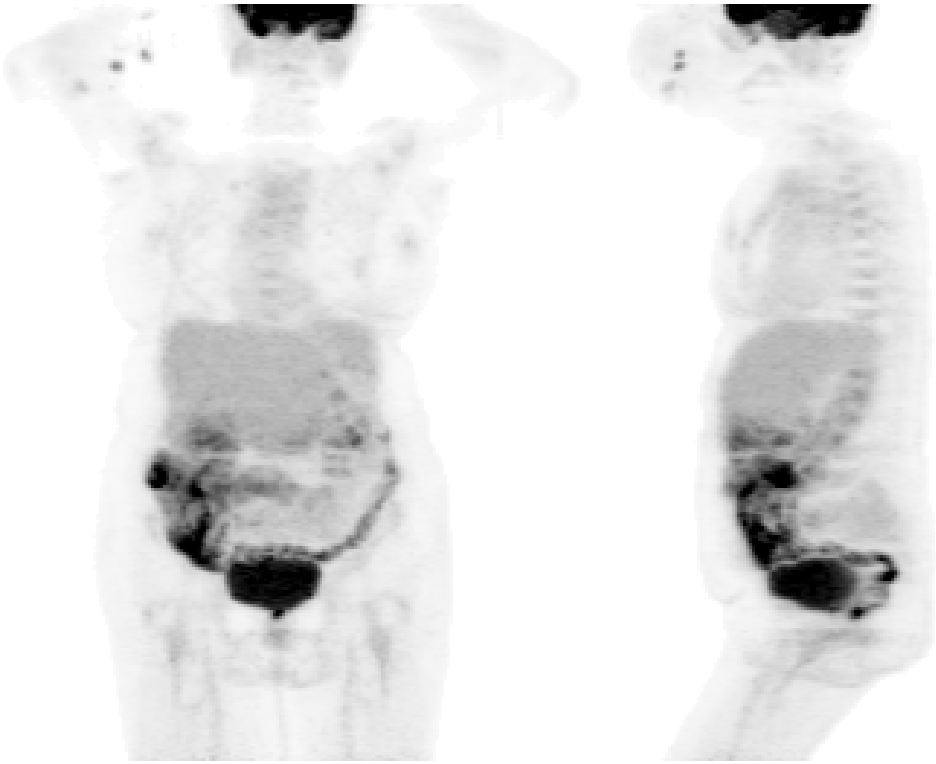


(a) Coronal view **(b)** Sagittal view

Figure 4.5: Dual-view PET MIPs of Patient 394559. Focal regions of asymmetrical tracer accumulation (potentially inflammatory or benign brown fat) visually masquerade as active lymphoma, resulting in false-positive predictions by the earlier network variants.

Driven by these suspicious visual artifacts, the Vision-Only model generates a false-positive score of 0.4504. The `BioMedText_Base` model, despite reading the text, barely crosses its threshold with a false-positive score of 0.3689. The breakthrough occurs exclusively with the `Hybrid_Attention` model. The tabular pathway mathematically isolates the continuous variable *"Time since initial diagnosis: 3.6 years"*.

In the clinical context of Hodgkin's lymphoma, a 3.6-year progression-free interval establishes an exceedingly low pre-test probability of recurrence. By explicitly gating the visual features with this long-term remission numerical prior, the hybrid architecture confidently suppresses the false alarm, outputting a highly accurate true-negative score of 0.2249.
The most extreme demonstration of tabular gating, however, is observed in the presence of severe clinical confounders, such as in Patient 403734 (True Label = 0). As evidenced in **Figure 4.6**, the patient's scan presents an overwhelming, systemic metabolic flare.



(a) Coronal view **(b)** Sagittal view

Figure 4.6: Dual-view PET MIPs of Patient 403734. The images reveal massive, diffuse radiotracer uptake throughout the axial skeleton and spleen, a classic artifact induced by recent G-CSF medication. This overwhelmingly deceptive visual signal caused both baseline models to fail.

The visual intensity of the bone marrow and splenic uptake is overwhelmingly deceptive, leading the Vision-Only model to a strong false-positive prediction of 0.4919. The `BioMedText_Base` model also fails dramatically (0.4516), unable to mathematically override the overwhelming visual signal simply by concatenating the text string *"underwent chemotherapy and G-CSF medication"*. In this extreme scenario, the `Hybrid_Attention` model leverages two critical numerical and categorical outliers: an extreme age (80 years old) and a massive temporal interval (*13.5 years* since initial diagnosis of hepatocellular carcinoma), explicitly paired with the G-CSF tabular flag. The cross-attention mechanism utilizes these structural features as a hard mathematical gate. Recognizing that a 13.5-year interval drastically reduces the

likelihood of original disease activity, and that the explicit G-CSF flag fully explains the severe skeletal uptake, the hybrid network effectively "shuts down" the visual false alarm, yielding a definitive true-negative probability of 0.2370.
Ultimately, these 34 critical rescue cases underscore a profound architectural insight: while language models are adept at capturing broad semantic contexts, the explicit tabular cross-attention mechanism is indispensable for enforcing precise numerical boundaries. This structural enhancement directly translates into the superior specificity and clinical reliability previously quantified on the independent test set.

5

Discussion and Conclusion

5.1 Summary of Main Findings

This research embarks on a fundamental challenge in computational oncology: transcending the intrinsic limitations of vision-only neural networks in the interpretation of complex functional imaging. By systematically designing, optimizing, and validating the Multimodal Lymphoma Artificial Intelligence Retrieval System (MM-LARS), we demonstrate that accurate ^{18}F-FDG PET/CT diagnosis requires the holistic orchestration of spatial, semantic, and numerical modalities.

Our investigation yields several critical insights, beginning with the optimization of the foundational single-modality baselines. We empirically prove that generic VLMs (such as PMC-CLIP and BioMedCLIP) suffer from a profound domain shift when applied to functional metabolic imaging, heavily underperforming compared to our domain-specific `LARS_ResNet` backbone. Furthermore, spatial ablation studies confirm that integrating orthogonal projections (coronal and sagittal) is indispensable, providing the network with a pseudo-3D spatial awareness necessary for disambiguating overlapping physiological radiotracer uptake. In parallel, our prompt engineering experiments quantify the impact of semantic granularity. By progressively enriching the textual input up to `prompt_v4`, we establish that explicit clinical histories—specifically recent treatments like chemotherapy or G-CSF—act as crucial cognitive constraints, effectively teaching the model to suppress visually deceiving inflammatory artifacts.

The most profound finding of this study, however, emerged during the architectural evolution of the multimodal fusion strategies. We discover that unstructured clinical narratives and structured numerical metadata interact with deep visual features in fundamentally different ways. While the early concatenation of text provides a stable baseline improvement, the naive injection of highly variable tabular data (e.g., patient age, exact time since diagnosis) into the `Hybrid_Fusion` model triggers a catastrophic collapse in diagnostic specificity, causing the network to overfit to metadata noise.

This optimization bottleneck is definitively resolved by our proposed `LARS_ResNet_Hybrid_Attentio` architecture. By explicitly utilizing the clinical text embedding as a continuous query to gate the tabular and visual features, the cross-attention mechanism successfully filters out numerical interference while amplifying critical temporal and demographic outliers. Evaluated on an independent, unseen test set, this tri-modal attention network not only achieves a high overall discriminative capability (AUC of 0.9403) but, more importantly, establishes a clinically optimal balance between sensitivity and

specificity. As further corroborated by our qualitative analysis, the hybrid network effectively mimics the diagnostic reasoning of expert nuclear medicine physicians—mathematically "explaining away" false-positive metabolic flares while confidently rescuing subtle false-negative visual presentations.

5.2 Limitations of the Current Study

Despite the promising diagnostic performance and the unprecedented clinical interpretability demonstrated by the MM-LARS framework, this study possesses several inherent limitations that warrant objective acknowledgment and future investigation.

The primary constraint of this research lies in its retrospective, single-center nature. Although the framework achieves an exceptional balance of sensitivity and specificity (AUC of 0.9403) on the independent test set, all ^{18}F-FDG PET/CT scans and electronic health records (EHR) are sourced from a single clinical institution. In real-world radiological practices, there exists substantial heterogeneity in imaging protocols across different hospitals, including variations in PET scanner vendors, radiotracer uptake times, and image reconstruction algorithms. Furthermore, the phrasing conventions of clinical prompts and the completeness of tabular metadata may fluctuate significantly across different healthcare systems. Consequently, whether the highly optimized cross-attention weights learned by the `Hybrid_Attention` model can seamlessly generalize to a multi- center, multi-vendor cohort without requiring domain adaptation remains to be prospectively validated.

A second limitation pertains to the model's reliance on the structural integrity of the clinical metadata. As systematically proven in the prompt engineering ablation study (Section 4.2.3), the diagnostic superiority of the multimodal network is heavily contingent upon the rich semantic granularity of `prompt_v4`, which explicitly includes critical variables such as recent chemotherapy or G-CSF administration. However, in routine clinical workflows, patient histories are frequently fragmented, improperly documented, or omitted entirely by referring physicians. While our dynamic serialization pipeline incorporates a graceful fallback mechanism to handle missing values (NaNs), a severe degradation in input metadata quality could inevitably restrict the text-guided attention mechanism, potentially causing the system's performance to regress toward the baseline visual-only network. Developing a more resilient cross-modal imputation strategy for highly sparse clinical scenarios is an ongoing necessity.

Finally, there is an inherent limitation regarding the granularity of visual interpretability. Section 4.5 successfully demonstrates that the tabular and textual attention pathways align closely with human clinical reasoning, effectively acting as semantic gates to rescue ambiguous false-negative and false-positive cases. However, this interpretability is predominantly evaluated at the decision-level (i.e., understanding why the model shifted its final probability score). The current framework lacks explicit pixel-level visual attribution. For optimal clinical trust, oncologists require algorithms that not only explain their reasoning textually but also precisely delineate the suspected hypermetabolic lesions directly on the PET MIPs (e.g., via class activation mapping or bounding boxes). The absence of such localized vi-

sual feedback limits the model's utility as a comprehensive, standalone diagnostic assistant.

5.3 Future Directions

Addressing the aforementioned limitations naturally paves the way for several promising avenues of future research, aiming to transition the MM-LARS framework from a highly controlled experimental environment to broad clinical deployment.
To ensure true universal generalizability, the immediate next step involves conducting multi-institutional, cross-vendor validation. Future iterations of this research should explore federated learning paradigms. By training the multimodal architecture across diverse hospital networks without centralizing raw patient data, the model can learn invariant metabolic representations that are robust against variations in PET scanner hardware, radiotracer dosage protocols, and local reconstruction algorithms.
From a data engineering perspective, reducing the framework's reliance on manually standardized prompts is a critical priority. Future developments could integrate large language models (LLMs) as dynamic upstream processing agents. Instead of relying on rigid templates, an integrated LLM could autonomously parse messy, unstructured EHR—extracting only the relevant histological subtypes, temporal intervals, and treatment histories—to generate a denoised, highly standardized `prompt_v4` embedding in real-time. This would significantly enhance the system's resilience against the fragmented documentation typically encountered in busy clinical workflows.
Lastly, bridging the epistemological gap between decision-level reasoning and spatial visualization remains a frontier for interpretability. While our tabular cross-attention mechanism successfully explains *why* a diagnosis was made, future architectures must seamlessly integrate pixel-level attribution. Incorporating weakly-supervised lesion segmentation branches or advanced gradient- weighted class activation mapping (Grad-CAM) directly conditioned on the text embeddings would allow the model to dynamically project bounding boxes or heatmaps onto the PET MIPs. Providing oncologists with both a text-guided probability score and a precise spatial localization of the hypermetabolic target would culminate in a truly transparent and interactive diagnostic assistant.

5.4 Conclusion

In conclusion, this thesis successfully develops and validates MM-LARS, a novel trimodal deep learning framework tailored for the complex diagnostic environment of lymphoma ^{18}F-FDG PET/CT. By systematically demonstrating that pure convolutional networks are inherently susceptible to visual ambiguities and physiological confounders, this research highlights the absolute necessity of contextual clinical priors. The proposed `Hybrid_Attention` architecture effectively bridges this gap, employing a sophisticated cross-attention mechanism to seamlessly orchestrate domain-specific visual features, semantically rich clinical narratives, and explicit numerical

tabular gates. Through rigorous quantitative evaluations and in-depth qualitative case studies, the framework proves its capacity to mimic human-like diagnostic reasoning, successfully rescuing evasive false negatives and suppressing deceptive false positives. Ultimately, MM-LARS represents a significant paradigm shift in computational oncology, moving beyond isolated image analysis toward a more holistic, robust, and clinically interpretable multimodal diagnostic standard.

Bibliography

[1] RL Siegel, KD Miller, NS Wagle, and A Jemal. Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, 73(1):17–48, 2023.

[2] OECD/European Commission. EU Country Cancer Profiles Synthesis Report 2025. Technical report, OECD Publishing, Paris, 2025.

[3] Jasmin Mettler, Horst Müller, Conrad-Amadeus Voltin, Christian Baues, Bernd Klaeser, Alden Moccia, Peter Borchmann, Andreas Engert, Georg Kuhnert, Alexander Drzezga, et al. Metabolic tumor volume for response prediction in advanced-stage hodgkin lymphoma. *Journal of nuclear medicine*, 60(2):207–211, 2019.

[4] R Frood, C Burton, C Tsoumpas, AF Frangi, F Gleeson, C Patel, and A Scarsbrook. Baseline pet/ct imaging parameters for prediction of treatment outcome in hodgkin and diffuse large b cell lymphoma: a systematic review. *European journal of nuclear medicine and molecular imaging*, 48(10):3198–3220, 2021.

[5] BD Cheson. Pet/ct in lymphoma: current overview and future directions. *Seminars in Nuclear Medicine*, 48(1):76–81, 2018.

[6] Cedric Rossi, Marie Tosolini, Pauline Gravelle, Sarah Pericart, Salim Kanoun, Solene Evrard, Julia Gilhodes, Don-Marc Franchini, Nadia Amara, Charlotte Syrykh, et al. Baseline suvmax is related to tumor cell proliferation and patient outcome in follicular lymphoma. *Haematologica*, 107(1):221, 2020.

[7] Robert Alexander, Stephen Waite, Michael A Bruno, Elizabeth A Krupinski, Leonard Berlin, Stephen Macknik, and Susana Martinez-Conde. Mandating limits on workload, duty, and speed in radiology. *Radiology*, 304(2):274–282, 2022.

[8] Cláudia S Constantino, Sónia Leocádio, Francisco PM Oliveira, Mariana Silva, Carla Oliveira, Joana C Castanheira, Ângelo Silva, Sofia Vaz, Ricardo Teixeira, Manuel Neves, et al. Evaluation of semiautomatic and deep learning–based fully automatic segmentation methods on [18f] fdg pet/ct images from patients with lymphoma: influence on tumor characterization. *Journal of Digital Imaging*, 36(4):1864–1876, 2023.

[9] Paul Blanc-Durand, Simon Jégou, Salim Kanoun, Alina Berriolo-Riedinger, Caroline Bodet-Milin, Françoise Kraeber-Bodéré, Thomas Carlier, Steven Le Gouill, René-Olivier Casasnovas, Michel Meignan, et al. Fully automatic segmentation of diffuse large b cell lymphoma lesions on 3d fdg-pet/ct for total metabolic tumour volume prediction using a convolutional neural network. *European Journal of Nuclear Medicine and Molecular Imaging*, 48(5):1362–1370, 2021.

[10] M Sadik, M Larsson, O Enqvist, L Edenbrandt, and E Trägårdh. Validation of an ai method for automated lymphoma metabolic tumor volume segmentation using a public benchmark pet/ct dataset. *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine*, pages jnumed–125, 2026.

[11] Ida Häggström, Doris Leithner, Jennifer Alvén, Gabriele Campanella, Murad Abusamra, Honglei Zhang, Shalini Chhabra, Lucian Beer, Alexander Haug, Gilles Salles, et al. Deep learning for [18f] fluorodeoxyglucose-pet-ct classification in patients with lymphoma: a dual-centre retrospective analysis. *The Lancet Digital Health*, 6(2):e114–e125, 2024.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[13] Ji Seung Ryu, Hyunyoung Kang, Yuseong Chu, and Sejung Yang. Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomedical Engineering Letters*, 15(5):809–830, 2025.

[14] Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A survey. *Medical Image Analysis*, 102:103551, 2025.

[15] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3):6, 2023.

[16] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.

[17] Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. Me-llama: Foundation large language models for medical applications. *Research square*, pages rs–3, 2024.

[18] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

[19] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15016–15027, 2023.

[20] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, 2022.

[21] Tong Lin, Jason Yan, David Jurgens, and Sabina J Tomkins. Tab2text-a framework for deep learning with tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12925–12935, 2024.

[22] Haigen Hu, Chao Du, Qiu Guan, Qianwei Zhou, Plerre Vera, and Su Ruan. A background-based data enhancement method for lymphoma segmentation in 3d pet images. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1194–1196. IEEE, 2019.

[23] Haigen Hu, Leizhao Shen, Tongxue Zhou, Pierre Decazes, Pierre Vera, and Su Ruan. Lymphoma segmentation in pet images based on multi-view and conv3d fusion strategy. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1197–1200. IEEE, 2020.

[24] Meng Wang, Huiyan Jiang, Tianyu Shi, and Yu-dong Yao. Hd-rds-unet: Leveraging spatial-temporal correlation between the decoder feature maps for lymphoma segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1116–1127, 2021.

[25] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[26] Russ A Kuker, David Lehmkuhl, Deukwoo Kwon, Weizhao Zhao, Izidore S Lossos, Craig H Moskowitz, Juan Pablo Alderuccio, and Fei Yang. A deep learning-aided automated method for calculating metabolic tumor volume in diffuse large b-cell lymphoma. *Cancers*, 14(21):5221, 2022.

[27] AS Cottereau, S Becker, F Broussais, O Casasnovas, S Kanoun, M Roques, N Charrier, S Bertrand, R Delarue, Christophe Bonnet, et al. Prognostic value of baseline total metabolic tumor volume (tmtv0) measured on fdg-pet/ct in patients with peripheral t-cell lymphoma (ptcl). *Annals of Oncology*, 27(4):719–724, 2016.

[28] Ludovic Sibille, Robert Seifert, Nemanja Avramovic, Thomas Vehren, Bruce Spottiswoode, Sven Zuehlsdorff, and Michael Schäfers. 18f-fdg pet/ct uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. *Radiology*, 294(2):445–452, 2020.

[29] Kwai Han Yoo. Staging and response assessment of lymphoma: a brief review of the lugano classification and the role of fdg-pet/ct. *Blood research*, 57(S1):75–78, 2022.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[31] Duoduo Li, Ruikun Li, Yutong Ma, Shengyun Huang, Miaofei Han, Yaozong Gao, and Ying Liang. Multi-view deep learning for automated lymphoma staging from 18f-fdg pet/ct: physician-level accuracy with high-throughput workflow. *EJNMMI research*, 2026.

[32] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):136, 2020.

[33] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[34] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[35] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

[36] Pengyu Wang, Huaqi Zhang, and Yixuan Yuan. Mcpl: multi-modal collaborative prompt learning for medical vision-language model. *IEEE Transactions on Medical Imaging*, 43(12):4224–4235, 2024.

[37] Sally F Barrington, N George Mikhaeel, Lale Kostakoglu, Michel Meignan, Martin Hutchings, Stefan P Müeller, Lawrence H Schwartz, Emanuele Zucca, Richard I Fisher, Judith Trotman, et al. Role of imaging in the staging and response assessment of lymphoma: consensus of the international conference on malignant lymphomas imaging working group. *Journal of clinical oncology*, 32(27):3048–3058, 2014.

[38] Bruce D Cheson, Richard I Fisher, Sally F Barrington, Franco Cavalli, Lawrence H Schwartz, Emanuele Zucca, and T Andrew Lister. Recommendations for initial evaluation, staging, and response assessment of hodgkin and non-hodgkin lymphoma: the lugano classification. *Journal of clinical oncology*, 32(27):3059–3067, 2014.

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY