Vision-Language Modeling for Medical Image Analysis

Dissertation Proposal

Jin Kim

Center for Computer Vision & Imaging Biomarkers

Committee:
Dr. Matthew S. Brown
Dr. Dan Ruan
Dr. John M. Hoffman
Dr. Michael McNitt-Gray
Dr. Yong Jae Lee

Abstract

Accurate analysis and interpretation of chest X-ray images is crucial for diagnosing and monitoring various pulmonary diseases. However, generating medical reports from these images is often time-consuming, labor-intensive, and subject to inter-observer variability. Most medical imaging AI involves building segmentation and classification systems to detect and measure anatomical structures or abnormalities. Very little investigation has been done on generating medical reports automatically, which is ultimately the task that must be accomplished by radiologists. We propose an integrated approach combining three key innovations: (1) extending the SimpleMind cognitive AI framework with medical-specific language and vision models through a flexible registry system and dual-path modality classification, (2) developing a specialized Vision-Language Model (VLM) using the MIMIC-CXR dataset, implementing Parameter-Efficient Fine-Tuning (PEFT) through LoRA and Prefix Tuning, and employing a novel three-stage training pipeline combining autoregressive pre-training, contrastive learning, and supervised fine-tuning, and (3) enhancing the model through multi-perspective informatic strategies including DCNv4 and Differential Transformer architectures for improved feature detection, multi-resolution preprocessing combining frequency domain analysis and adaptive spatial enhancement, and a structured clinical reasoning approach using Chain-of-Thought (CoT) prompting with specialized radiological templates. The system will be evaluated using held-out portions of the MIMIC-CXR dataset through an enhanced framework combining GREEN methodology with semantic embedding-based assessment. Our hypothesis is that this comprehensive VLM approach, integrating sophisticated model architectures with efficient adaptation techniques and domain-specific enhancements, will result in a robust and efficient system for generating medical reports that outperforms current state-of-the-art methods in terms of accuracy, interpretability, and computational efficiency. This work makes technical contributions in applying VLMs to medical imaging, and has the potential to significantly improve the efficiency and consistency of chest X-ray interpretation with clinical accuracy.

Specific Aims

Accurate analysis and interpretation of medical images is crucial for diagnosis and patient care across various medical specialties. However, current approaches to medical image analysis often require substantial manual effort, leading to workflow inefficiencies and potential variability in interpretation. Most medical imaging AI involves building segmentation and classification systems using Convolutional Neural Networks (CNNs) to detect and measure anatomical structures or abnormalities. While these systems excel at specific tasks, they are fundamentally limited by their local feature processing nature and inability to provide comprehensive interpretations as in radiologist reports. Very little investigation has been done on generating medical reports from images automatically.

Vision-Language Models (VLMs) offer several crucial advantages over traditional CNN approaches for medical image analysis. First, their transformer-based architecture enables global context processing, allowing them to capture long-range relationships between anatomical structures and pathological findings immediately - a critical capability for holistic image interpretation. Second, VLMs excel at connecting visual information with semantic meaning, enabling them to not only detect abnormalities but also describe them in clinically relevant language. This multimodal understanding is essential for automated report generation that matches the depth and nuance of expert radiologist interpretations. Finally, VLMs demonstrate superior scaling capabilities when trained on large datasets, potentially allowing them to handle the wide variety of appearances and conditions encountered in clinical practice.

Recent advancements in medical-specific language models (Med-PaLM 2, BioGPT), VLMs (MedViLT, PMC-CLIP), and PEFT techniques offer new opportunities to address these challenges comprehensively. We hypothesize that by integrating these advanced AI capabilities through a carefully designed framework combining model registry systems, efficient fine-tuning strategies, and multi-perspective informatic approaches, we can create a more automated, reliable, and efficient system for medical image analysis and reporting that better aligns with actual clinical workflows. By leveraging the comprehensive MIMIC-CXR dataset throughout our research, we aim to validate these approaches while developing methods that can generalize to other medical imaging domains. This integrated approach has the potential to significantly improve medical image analysis workflows while maintaining the high standards of accuracy and interpretability required in clinical settings.

SA-1: Integrate Language and Vision Model agents into the SimpleMind Cognitive AI Framework

We will extend the SimpleMind cognitive AI framework by integrating state-of-the-art language and multimodal models. We will implement a model registry system that integrates, and enables switching between, popular open-source medical LLMs (Med-PaLM 2, BioGPT, ClinicalT5) and VLMs (MedViLT, PMC-CLIP, PathCLIP). This system will utilize flexible adapter agents to handle model-specific requirements such as tokenization and output formatting, while specialized fusion agents will combine SimpleMind's vision capabilities with medical LLM contextual understanding. We will test the integration by building two applications: (1) automated generation of SimpleMInd configurations from natural language using LLMs, (2) image classification and characterization using VLMs to direct (orchestrate) downstream processing. These applications are proof of concept only for integration testing, formal evaluation is outside the scope of the thesis.

SA-2: Develop a 2D Vision-Language Model for Chest X-ray Medical Report generation

We will develop a specialized VLM for generating accurate medical reports from chest X-ray images using the MIMIC-CXR dataset. We will use Low-Rank Adaptation (LoRA) techniques to tune large models to our specific application. We will adapt the LLaVA architecture, combining a CLIP ViT-L/14 vision encoder with a LLaMA-based language model. Our hypothesis is that this approach will achieve comparable performance than current state-of-the-art methods for medical report generation on MIMIC-CXR dataset (BLEU-4 > 0.133, ROUGE-L > 0.289, METEOR > 0.167, CIDEr > 0.241). We will evaluate the system using standard text generation metrics (BLEU, ROUGE-L, CIDEr) and clinical accuracy metrics (F1-score on critical findings). These metrics will serve as baselines for measuring the improvements achieved through the enhancements proposed in SA-3.

SA-3: Enhance the 2D Vision-Language Model with multi-perspective informatic strategies for Improved Report Generation

We will enhance the performance of our base VLM through a comprehensive multi-perspective approach focusing on four key technical innovations. First, we will optimize the image encoder by implementing and fine-tuning both the Deformable Convolution Network v4 (DCNv4) and Differential Transformer architectures. Second, we will develop a multi-resolution preprocessing framework that combines frequency domain analysis through wavelet

transforms and adaptive spatial resolution enhancement. Third, we will implement a structured clinical reasoning approach through CoT prompting enhanced with specialized radiological templates, enabling systematic analysis that mirrors expert diagnostic processes through multiple stages of increasing complexity. Finally, we will optimize the training process through a curriculum learning approach combined with knowledge distillation, organizing cases by complexity from normal PA views to complex conditions, while implementing a novel distillation protocol to transfer expertise to more compact models suitable for clinical deployment. We will evaluate these enhancements through an integrated framework combining extensions to the GREEN methodology with semantic embedding-based assessment, enabling partial credit for semantically related alternatives while maintaining strict evaluation for critical findings. The effectiveness of these innovations will be validated using a held-out portion of the MIMIC-CXR dataset, with particular attention to improvements in report accuracy, clinical relevance, and computational efficiency. We hypothesize that our enhancements will improve the base VLM from SA-2 by at least 10% in report accuracy, clinical relevance, and computational efficiency metrics. To enable direct comparison of our enhancements against the base VLM developed in SA-2, we will use the identical dataset splits: 368,960 training, 2,991 validation, and 5,159 test cases. This consistent dataset strategy ensures that improvements in metrics can be directly attributed to our architectural and methodological enhancements rather than differences in training data.

Background and Significance

The analysis and interpretation of medical images, particularly chest X-rays, remains a critical yet challenging task in healthcare delivery.[1,2] While VLMs have emerged as powerful tools for understanding and describing visual information,[3] the medical field faces significant challenges in implementing these technologies for clinical applications.[4] Current medical report generation processes are notably time-consuming and require extensive professional expertise, resulting in high operational costs for healthcare institutions.[5] The subjective nature of medical image interpretation can lead to significant inter-observer variability, potentially affecting diagnostic consistency and patient care quality.[6] These challenges are compounded by both the scarcity of high-quality annotated medical imaging data and the substantial computational resources demanded by traditional approaches to model training.[7] Traditional evaluation metrics for medical report generation include Bilingual Evaluation Understudy (BLEU)[8] for precision-based assessment, Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[9] for recall-based evaluation, and Consensus-based Image Description Evaluation (CIDEr)[10] for measuring consensus through TF-IDF weighted n-gram similarities, though each has limitations in capturing the nuances of medical terminology and clinical significance.

Traditional CNN-based approaches in medical imaging have primarily focused on specific diagnostic tasks like lesion detection, segmentation, and classification.[11,12] While effective for these targeted applications, CNNs are inherently limited by their architectural design. They process images through hierarchical local feature extraction, making them well-suited for detecting visual patterns but less capable of understanding broader contextual relationships or generating natural language descriptions.[13] This fundamental limitation creates a significant gap between AI outputs and the comprehensive reports that radiologists must produce.[14] In contrast, VLMs represent a paradigm shift in medical image analysis by combining visual understanding with natural language capabilities.[15] Their transformer-based architecture enables global context processing, allowing them to capture long-range relationships between anatomical structures and pathological findings. This ability to connect visual features with semantic meaning enables VLMs to not only detect abnormalities but also describe them in clinically relevant language, more closely matching the way radiologists actually work. Furthermore, VLMs can leverage knowledge from both visual and textual medical data during pre-training, potentially enabling better generalization across different medical conditions and imaging protocols.

Recent architectural advances in computer vision have opened new possibilities for addressing these challenges. The DCNv4[16] represents a significant breakthrough in image processing through its innovative dual-branch architecture, where dynamic sampling parameters are computed through specialized convolutional layers to adapt to varying anatomical structures. This adaptation capability is particularly crucial for medical imaging, where subtle variations can have significant diagnostic implications. Complementing this, the Differential Transformer[17] architecture has demonstrated remarkable capabilities in medical feature detection through its novel attention mechanism, which calculates attention scores as differences between separate softmax attention maps, enabling more precise distinction of pathological features from normal anatomy.

PEFT techniques have emerged as another crucial advancement in making sophisticated medical AI more accessible. LoRA[18] achieves remarkable efficiency by decomposing weight updates into low-rank matrices, enabling model adaptation with updates to only 0.1-4% of parameters. Prefix Tuning[19] complements this by prepending trainable continuous prompts while keeping the base model frozen, preserving model knowledge while enabling task-specific adaptation. These approaches are particularly valuable in medical settings where computational resources may be limited but model performance cannot be compromised.

The integration of clinical reasoning patterns into AI systems has seen significant advancement through CoT[20,21] prompting frameworks. These frameworks enable systematic analysis that mirrors expert diagnostic processes, breaking down complex interpretations into explicit reasoning steps that enhance both accuracy and interpretability. When combined with curriculum learning[22] strategies that progressively increase complexity from normal anatomical structures to complex pathological findings, these approaches show promise in developing more clinically relevant analysis systems. The addition of multi-resolution preprocessing techniques, incorporating both frequency domain analysis through wavelet transforms[23] and adaptive spatial enhancement, enables comprehensive capture of both fine-grained pathological details and broader anatomical context.

The significance of our proposed research lies in its comprehensive integration of these advances into a practical clinical system. First, by extending the SimpleMind cognitive AI framework[24,25] with a flexible model registry system and dual-path modality classification, we enable seamless integration of advanced AI capabilities while

maintaining clinical standards through RadLex[26]-based terminology validation. Second, our PEFT approach combining LoRA and Prefix Tuning makes sophisticated medical image analysis more accessible to institutions with limited computational resources, potentially reducing healthcare disparities in access to advanced diagnostic tools. Third, our multi-perspective informatic strategies, including DCNv4 and Differential Transformer architectures, multi-resolution preprocessing, and structured clinical reasoning, address the complex challenges of medical image interpretation while maintaining high standards of accuracy and interpretability.

Beyond immediate clinical applications, this research offers broader implications for healthcare delivery:

Enhanced Clinical Workflow Integration Our systems will be implemented in the SimpleMind AI platform. SimpleMind utilizes a blackboard architecture that facilitates communication between traditional image processing components and new AI capabilities. It has been integrated unti clinical imaging workflows at UCLA. Thus, our strategy ensures that advanced AI technologies enhance rather than disrupt established clinical practices.

Democratized Access to Advanced AI By implementing efficient fine-tuning strategies and knowledge distillation techniques that achieve 4x model compression while maintaining 95% performance, our system makes advanced medical image analysis more accessible to healthcare institutions with varying resource levels. This democratization could help reduce healthcare disparities by enabling wider deployment of sophisticated diagnostic support tools.

Improved Clinical Standardization Our structured clinical reasoning approach, combined with curriculum learning and systematic evaluation frameworks, promotes more consistent medical interpretations while maintaining the flexibility to handle diverse radiological cases. This standardization could help reduce inter-observer variability while preserving the crucial role of clinical expertise in final decision-making.

This research addresses critical gaps in current medical imaging workflows by combining cutting-edge AI technologies with practical clinical considerations. By integrating sophisticated model architectures, efficient adaptation techniques, and domain-specific enhancements within a comprehensive framework, we aim to significantly improve the efficiency and consistency of medical image interpretation while maintaining high standards of clinical accuracy and interpretability. The potential impact extends beyond immediate technical advancements to broader improvements in healthcare delivery through enhanced efficiency, accessibility, and standardization of medical image analysis.

Research Design and Methods

SA-1 : Integrate Language and Vision Model agents into the SimpleMind Cognitive AI Framework

SimpleMind Framework Integration Methodology

We will extend the SimpleMind (SM) cognitive AI framework to seamlessly integrate state-of-the-art language and multimodal models while leveraging its existing computer vision capabilities. Building upon SimpleMind's blackboard architecture, we will implement a model registry system that enables flexible integration and switching between different open-source medical LLMs (including Med-PaLM 2,[27] BioGPT,[28] and ClinicalT5[29]) and VLMs (such as MedViLT,[30] PMC-CLIP,[31] and PathCLIP[32]). The knowledge base will focus on RadLex-based terminology validation and uncertainty quantification to ensure generated reports maintain clinical standards while appropriately expressing confidence levels. The core of our contribution lies in developing flexible adapter agents that enable efficient integration of these pre-trained models within the SimpleMind ecosystem. These adapters will handle model-specific requirements such as tokenization, prompt engineering, and output formatting. For cross-modal operations, we will implement specialized fusion agents that combine SimpleMind's established vision capabilities with contextual understanding from medical LLMs. This design ensures we can rapidly incorporate new models as they become available while maintaining consistent interfaces and validation standards. To coordinate these models effectively, we will optimize the blackboard architecture for efficient interaction between SimpleMind's vision agents and the newly integrated language and multimodal capabilities. The system will employ a streamlined validation pipeline that standardizes terminology through RadLex mapping and appropriately quantifies uncertainty in generated descriptions. This practical approach enables reliable integration of advanced AI capabilities while maintaining SimpleMind's established strengths in medical image analysis.

Integration test application: SimpleMind configuration using LLMs

The purpose of this integration test is to demonstrate that LLMs can successfully understand natural language descriptions of medical image analysis tasks and convert them into valid SimpleMind configurations. This capabil-

ity would significantly streamline the process of setting up new SM applications, reducing the technical expertise required and making the framework more accessible to clinical users. SM applications are configured by linking processing agents in a graph that represents dependencies between agent inputs and outputs. This graph of agents is specified in a structured YAML format that is used to conifugre the system. We will implement and evaluate the integration of Large Language Models (LLMs) within the SimpleMind framework, focusing on automated YAML configuration generation from natural language prompts. As illustrated in Figure 2, our system implements a streamlined pipeline with three main components: natural language processing, configuration generation, and validation. The natural language processing component analyzes user inputs to identify key task elements, including the processing type (e.g., segmentation), target organs, and required processing steps. These structured requirements are then processed by a medical-domain LLM that has been fine-tuned with SimpleMind-specific knowledge, enabling it to generate appropriate YAML configurations. A validation layer ensures both syntactic correctness and semantic consistency by verifying agent dependencies and parameter constraints before producing the final configuration.

The implementation will begin with developing a standardized interface for LLM agents that can accommodate different model architectures, ensuring flexibility in model selection and future upgrades. We will create a specialized prompt engineering system that converts SimpleMind's knowledge base requirements into structured natural language queries, enabling the automatic generation of application configurations through LLM interpretation and response parsing. The framework will be designed to process VLM outputs, which typically consist of detailed image descriptions including modality identification, anatomical structure recognition, and pathological finding descriptions. We will develop parsing agents that can transform these descriptive outputs into structured data compatible with SimpleMind's blackboard architecture, enabling integration with existing image processing and reasoning agents. The system will implement both rule-based and learned approaches for extracting relevant information from VLM outputs, with particular attention to maintaining the semantic relationships crucial for medical image understanding. Other approaches we have considered include end-to-end neural architectures and template-based parsing systems, but we will focus on developing a hybrid approach that combines the flexibility of language models with the reliability of structured knowledge representation. The feasibility testing will particularly focus on the automatic generation of SimpleMind configurations from natural language descriptions

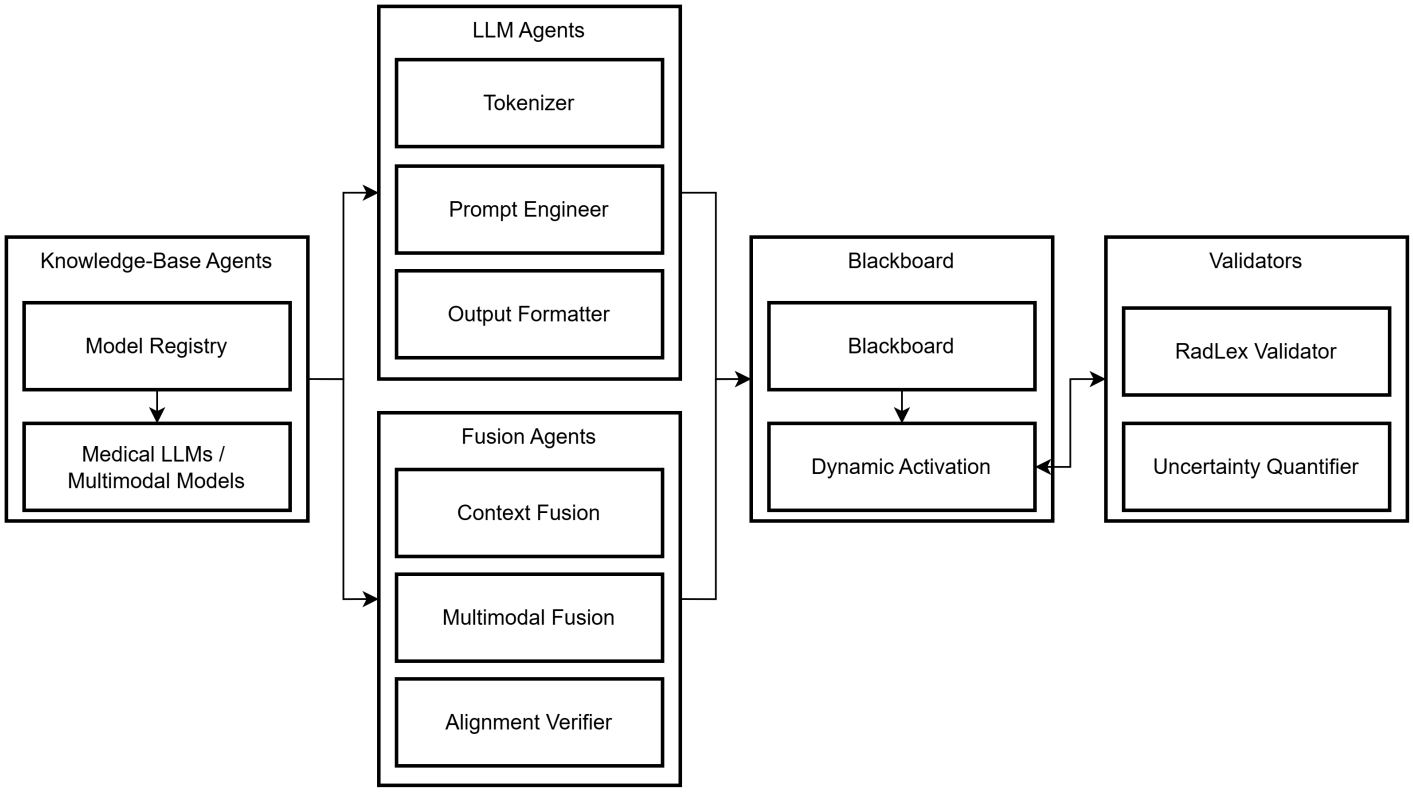


Figure 1. Proposed SimpleMind framework extension for language and multimodal model integration. The framework consists of four main components: (1) Knowledge-Base Agents managing model registry and pre-trained models, (2) LLM and Fusion Agents handling model-specific adaptations and multimodal fusion, (3) Blackboard system coordinating agent interactions, and (4) Validators ensuring medical accuracy through RadLex terminology validation and uncertainty quantification. This modular design enables flexible integration of new models while maintaining clinical standards.

of medical image analysis tasks. This will include developing evaluation metrics for configuration quality, testing the system's ability to handle various medical imaging scenarios, and validating the generated configurations against expert-created ones. Through this systematic approach, we aim to demonstrate the feasibility of using LLM agents to streamline SimpleMind application development while maintaining the framework's core strengths in medical image analysis.

Integration test application: Image classification and characterization using VLMs

The purpose of this integration test is to verify that VLMs can effectively analyze medical images and provide high-level guidance for SimpleMind's processing pipeline, specifically by providing a structured report on an input image modality parameters and anatomic coverage so that appropriate processing can be triggered. By using VLMs to identify image modalities and key characteristics, we can automatically select and configure appropriate downstream processing modules ("AI orchestration"). We will implement and evaluate a dual-path modality classification system as the initial gateway for SimpleMind pipeline orchestration, leveraging both VLMs and traditional deep learning approaches (Figure 3). The primary path utilizes VLMs to interpret medical images through natural language understanding, while maintaining an efficient CNN-based alternative path for comparison and fallback purposes. The VLM-based approach capitalizes on these models' sophisticated visual understanding capabilities and their ability to process and describe image characteristics in natural language. We will develop specific prompting strategies (e.g., "Describe the imaging modality and key characteristics of this medical image") and robust output parsing mechanisms to extract modality classifications from VLM outputs. To enable accurate modality and parameter detection, we will create a curated reference set where each entry pairs an image with a structured modality report. This report will capture hierarchical information including: (1) Primary classification details (modality type, sub-modality variants), (2) Technical parameters (radiation attributes, image quality metrics, artifact patterns), (3) Tissue visualization properties (density differentiation, contrast characteristics), (4) Anatomical coverage specifications, and (5) Clinical quality indicators. For example, the report includes cross-sectional properties and reconstruction parameters for CT, while X-ray entries detail projection properties and structure superposition characteristics.

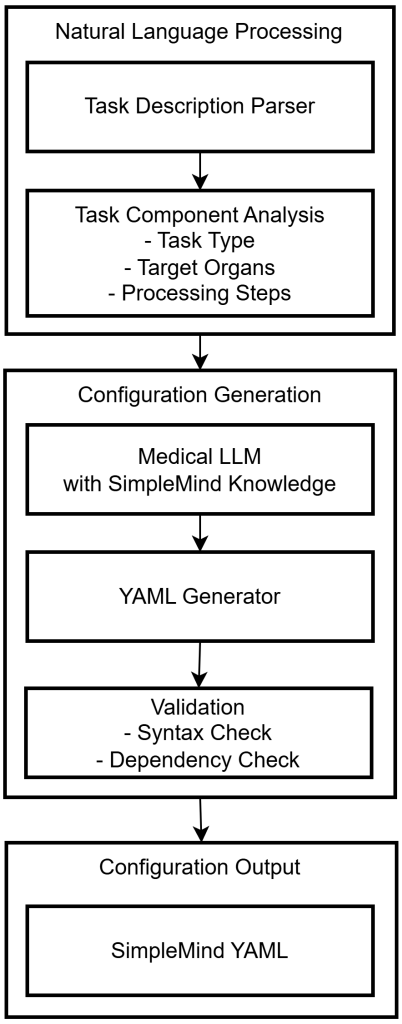


Figure 2. SimpleMind configuration generation architecture showing the three main components: (1) Natural Language Processing for task description parsing and component analysis, (2) Configuration Generation utilizing a medical LLM with SimpleMind-specific knowledge for YAML generation, and (3) Configuration Output with validation checks for syntax and dependencies. This streamlined pipeline enables automatic generation of valid SimpleMind configurations from natural language descriptions.

To enable accurate classification and parameter detection, we will create a curated reference dataset where each entry contains both an image and its corresponding structured modality report. This comprehensive reporting framework allows the VLM to not only identify the basic modality type but also extract crucial technical parameters that influence downstream processing steps. For example, when analyzing an MRI scan, the system will determine specific sequence types (T1/T2-weighted), field strength, slice characteristics, and tissue contrast patterns - information vital for subsequent image processing and analysis tasks. The modality reports will be designed to capture both the visual characteristics that VLMs excel at describing and the technical parameters essential for medical image processing pipelines. Generation of these structured modality reports will be of value for DICOM images, where header information is often incomplete, and especially for images in other formats, such as nifti, which have only very limited meta information.

This approach aligns with the broader trend of leveraging large language models in medical imaging while potentially offering more nuanced understanding of complex cases. As a complementary approach, we maintain a traditional classification path using lightweight convolutional neural networks (e.g., ResNet-18, MobileNet) that have demonstrated strong performance in medical image classification tasks. This path implements standard

preprocessing steps including intensity normalization and size standardization, offering computational efficiency and proven reliability for basic modality classification.

Although formal evaluation is outside the scope of this integration test, we will assess both approaches across common medical imaging modalities (X-ray, CT, MRI, ultrasound), using a comprehensive dataset that includes diverse examples with varying view angles, contrast levels, and image quality. Performance metrics will track classification accuracy, processing time, and reliability, with particular attention to each approach's strengths in different scenarios. Quality assurance mechanisms include confidence score monitoring, confusion matrix analysis, and validation against expert-labeled test sets. Through this dual-path strategy, we aim to leverage the sophisticated understanding capabilities of VLMs while maintaining the option of proven CNN-based classification methods. This approach provides flexibility in deployment scenarios while ensuring reliable modality classification for SimpleMind's pipeline orchestration.

SA-2 : Develop a 2D Vision-Language Model for Chest X-ray Medical Report Generation

Dataset Curation and Preprocessing

For developing our medical report generation system, we plan to utilize the MIMIC-CXR dataset, a large-scale chest X-ray dataset with paired radiological reports. The dataset contains 377,110 chest x-ray images corresponding to 227,835 radiographic studies, split into 368,960 training, 2,991 validation, and 5,159 test cases, and detailed annotations for 14 different findings (including pathological conditions and support devices), labeled using both CheXpert and NegBio natural language processing tools. Figure 4 provides a comprehensive analysis of the dataset characteristics we will work with. The correlation matrix (Figure 4 (a) reveals several clinically relevant relationships between findings, with notable correlations between atelectasis and pleural effusion (0.31), edema and pleural effusion (0.25), and lung opacity and pneumonia (0.19). The prevalence of individual findings (Figure 4 (b) indicates that pleural effusion (23.8%), lung opacity (22.6%), and atelectasis (20.1%) are the most common observations, while fractures (1.9%) and other pleural conditions (0.9%) are relatively rare.

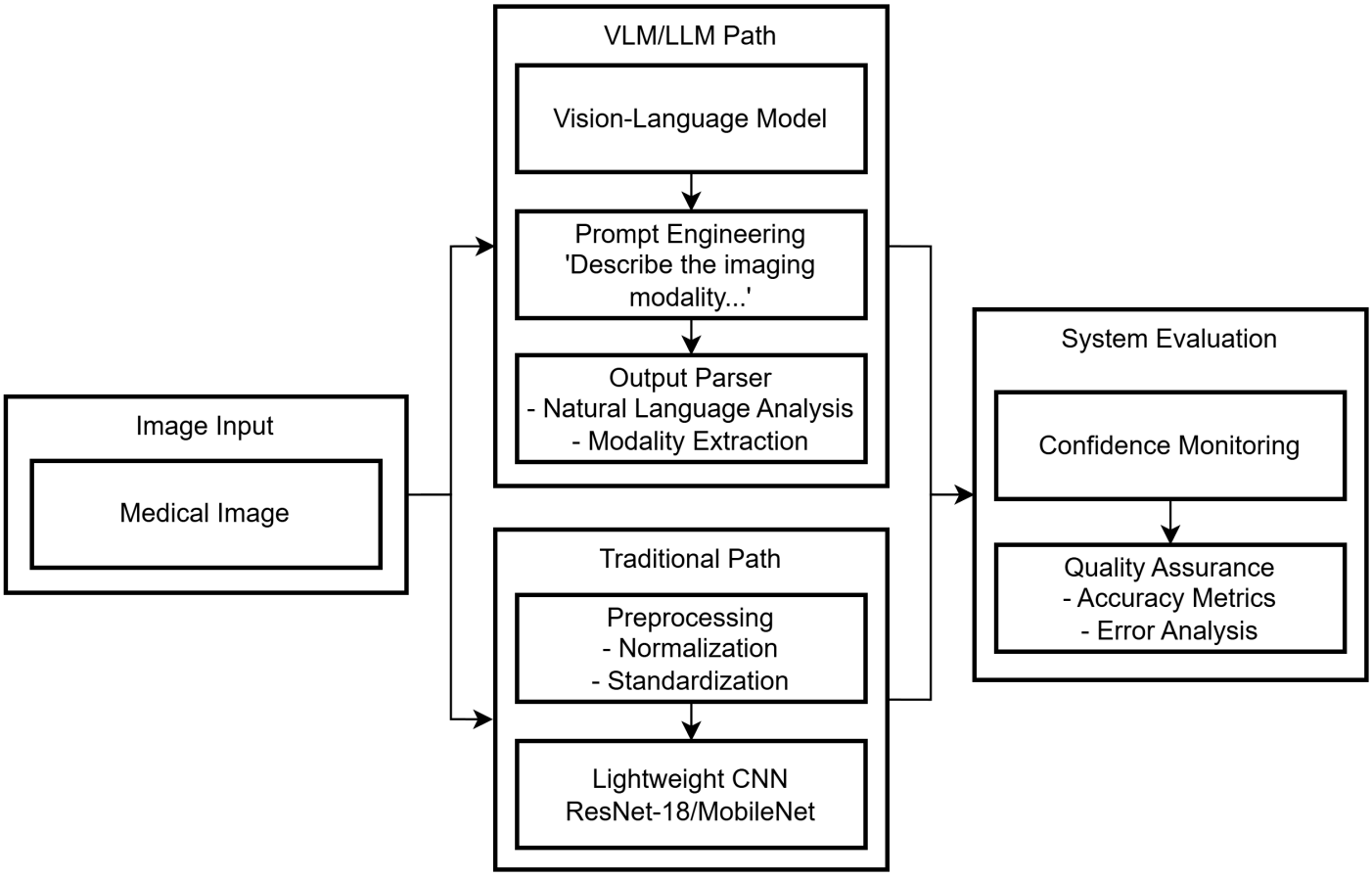


Figure 3. Dual-path medical image modality classification system architecture. The primary path (top) utilizes VLMs with natural language prompting and parsing for sophisticated image understanding. The alternative path (bottom) implements a lightweight CNN approach with traditional preprocessing for efficient classification. Both paths feed into a unified evaluation system that monitors confidence and ensures quality assurance.

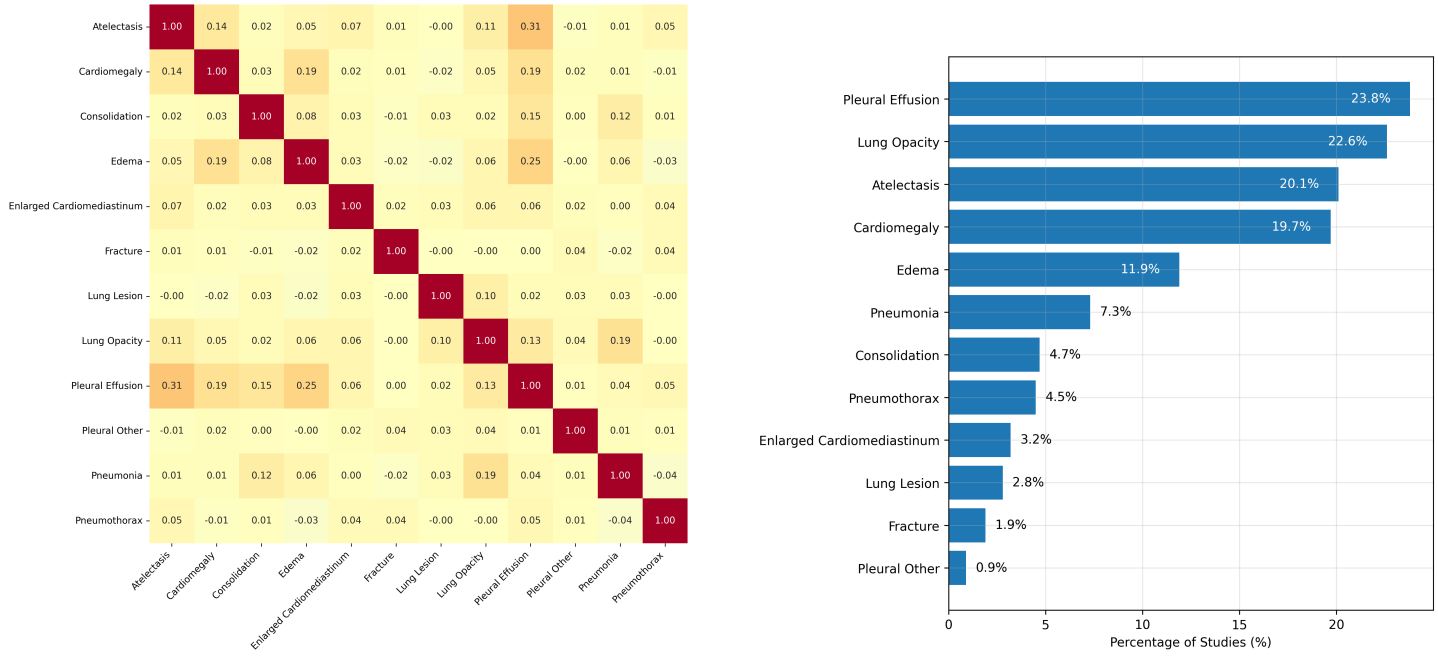
We will preprocess the dataset following the MIMIC-CXR-JPG conversion protocol. The chest X-ray images will be accessed in JPG format, which were converted from DICOM using a standardized process: pixel values normalized to the range [0, 255] by subtracting the lowest value, dividing by the highest value in the shifted image, and converting to unsigned integers. The images are already preprocessed with histogram equalization for contrast enhancement and stored with a quality factor of 95. For the reports, we will utilize the provided structured labels that were extracted from either the impression section (82.4% of reports), findings section (12.5%), or the final section (5.1%) when neither was present. These labels are classified as positive (1.0), negative (0.0), or uncertain (-1.0). Following the official dataset split strategy, we will divide the data into training, validation, and test sets while maintaining the distribution of findings across all partitions and preventing patient overlap between sets.

Parameter Efficient Fine Tuning

We will develop a PEFT approach that combines LoRA and Prefix Tuning to enable efficient adaptation of large VLMs for medical report generation. As shown in Figure 5, our implementation comprises two complementary strategies. LoRA (Figure 5 (a) decomposes weight updates into low-rank matrices, where for each pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, we learn a decomposition $\Delta W = BA$ with $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$. Prefix Tuning (Figure 5 (b)) prepends trainable continuous prompts to the input sequence while keeping the transformer layers frozen. This architecture allows us to optimize only a small number of parameters (0.1-4% of original model size) while maintaining model performance. We will implement these methods with particular focus on the cross-attention layers between visual and textual components, setting rank $r = 8$ for LoRA and prefix lengths $l_v = 10$ and $l_t = 5$ for visual and textual components respectively. Alternative approaches considered include adapter layers and selective fine-tuning, but we focus on the LoRA-Prefix combination for its superior parameter efficiency and ability to preserve model quality. Our hypothesis is that this combined approach will enable effective model adaptation while maintaining minimal memory requirements and computational overhead, making it practical for clinical deployment.

Model Architecture and Training

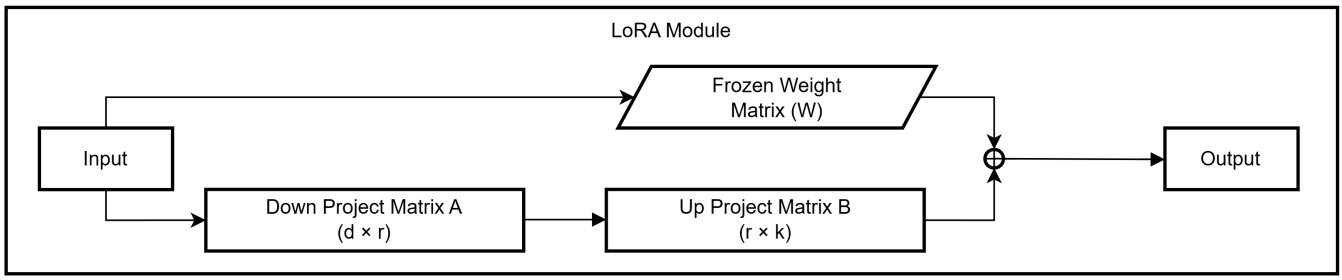
We will implement and optimize the LLaVA (Large Language and Vision Assistant) architecture as our baseline model for medical report generation. As shown in Figure 7, LLaVA combines a CLIP ViT-L/14 vision encoder with a LLaMA-based language model through a learnable linear projection layer, enabling effective cross-modal interaction for medical image understanding and report generation. The vision encoder processes X-ray images through a series of transformer blocks, generating visual embeddings that capture both local anatomical details and global structural patterns. These visual features are then projected into the language model's embedding



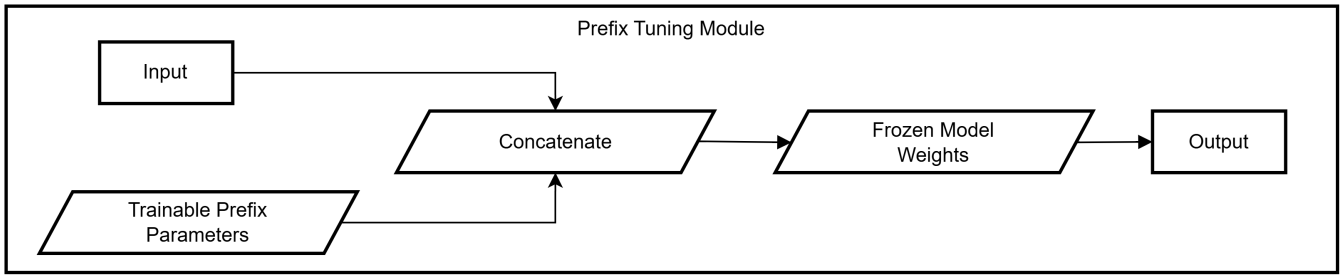
(a) Correlation matrix between different findings

(b) Prevalence of individual findings

Figure 4. Analysis of the MIMIC-CXR dataset characteristics showing (a) correlations between different findings (darker red indicates stronger correlation), and (b) prevalence of individual findings across the dataset.



(a) LoRA architecture



(b) Prefix Tuning architecture

Figure 5. PEFT strategies. (a) LoRA decomposes weight updates into two trainable low-rank matrices A and B ($\mathbb{R}^{d\times r}$ and $\mathbb{R}^{r\times k}$ respectively), enabling efficient adaptation while keeping pre-trained weights frozen. (b) Prefix Tuning prepends trainable continuous prompt vectors before the transformer layers, allowing the model to learn task-specific behaviors through prefix parameters while keeping the base model frozen. In both approaches, red components indicate trainable parameters while gray indicates frozen components.

space, where a decoder-only transformer architecture processes them alongside textual input to generate detailed medical reports through autoregressive generation. Our training strategy, illustrated in Figure 6, employs a novel three-stage approach for adapting the model to medical report generation. We begin by initializing the model with original LLaVA weights to establish a robust foundation for vision-language understanding. The autoregressive pre-training stage focuses on optimizing the vision encoder's ability to capture medical imaging features through patch-based processing and transformer encoding. This is followed by a contrastive learning phase that aligns visual and textual embeddings in a shared semantic space, crucial for accurate mapping between radiological findings and their descriptions. The final supervised fine-tuning stage specifically targets chest X-ray report generation using the full MIMIC-CXR dataset.

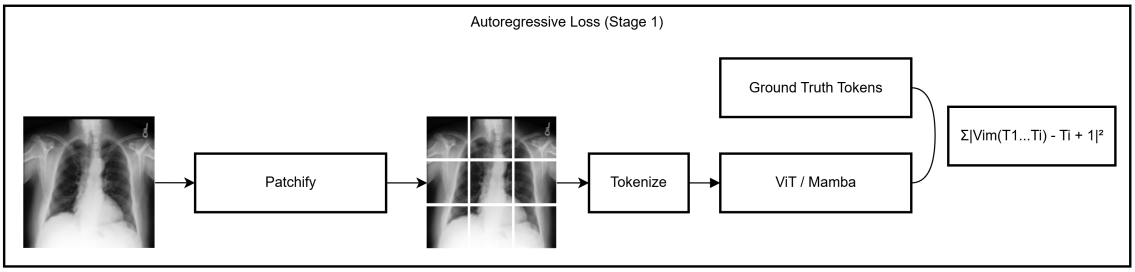
To ensure stable and efficient training, we implement several technical optimizations:

- Custom loss functions combining cross-entropy for text generation and medical term accuracy
- Gradient accumulation and dynamic learning rate scheduling
- Specialized cross-attention mechanisms optimized for radiological feature interpretation
- End-to-end differentiability while leveraging pre-trained weights

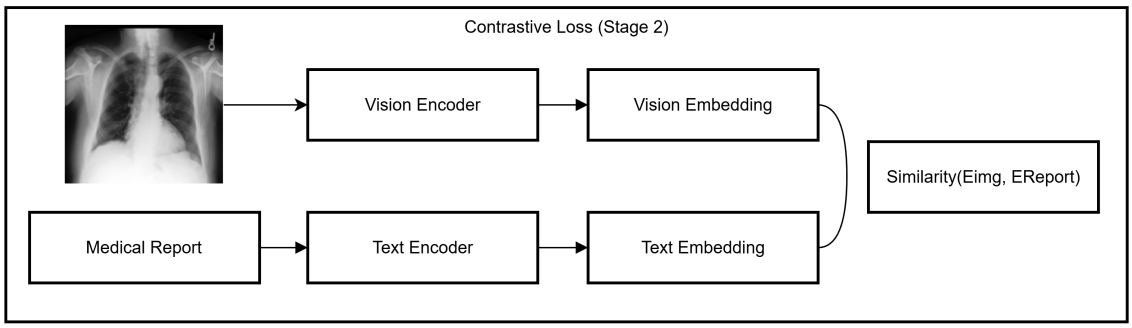
The training process is monitored through comprehensive metrics including perplexity on validation reports, standard text generation metrics (BLEU, ROUGE-L, CIDEr), and clinical accuracy metrics specific to medical findings. This implementation serves as a foundation for evaluating subsequent enhancements in SA-3, particularly the integration of multi-resolution capabilities and reasoning agents.

Evaluation methods

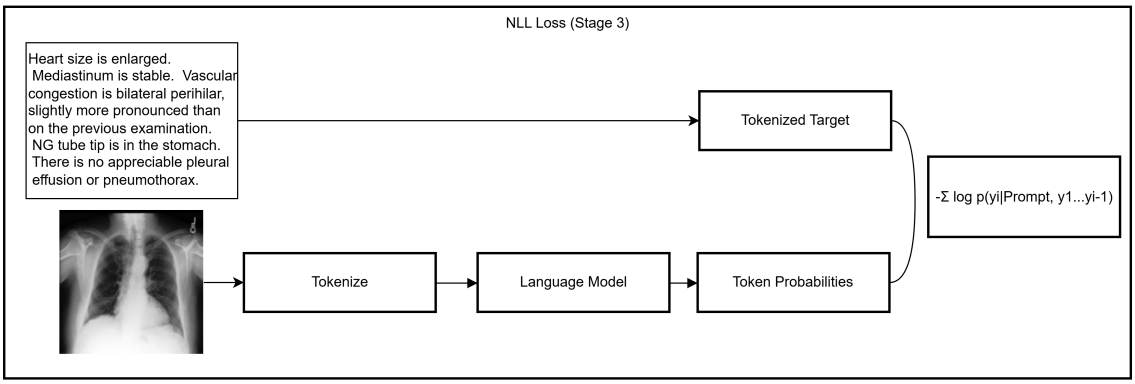
To assess the effectiveness of our medical report generation system, we implement a comprehensive evaluation framework utilizing standard natural language generation metrics while considering the unique challenges of medical text evaluation. Our evaluation approach addresses three key aspects: text generation quality, clinical accuracy, and radiological consistency. For text generation quality, we employ three complementary metrics that capture different aspects of report accuracy. BLEU evaluates the precision of n-gram matches between generated reports and reference texts. ROUGE-L measures the recall of the longest common subsequence between generated and reference reports. CIDEr assesses consensus by computing TF-IDF weighted n-gram similarities between the generated report and multiple references. Recent state-of-the-art methods have achieved performance metrics of BLEU-4 = 0.133, ROUGE-L = 0.289, METEOR = 0.167, and CIDEr = 0.241 on medical



(a) Stage 1: Autoregressive pre-training



(b) Stage 2: Contrastive learning



(c) Stage 3: Supervised fine-tuning

Figure 6. Three-stage training pipeline for medical report generation. (a) Autoregressive pre-training processes X-ray images through patchify operation and ViT/Mamba encoding, optimizing for accurate reconstruction of visual features through ground truth token prediction. (b) Contrastive learning phase trains vision and text encoders to map X-ray images and medical reports into a shared embedding space, optimizing similarity between corresponding image-report pairs. (c) Supervised fine-tuning uses NLL loss to train the language model for generating accurate medical reports by minimizing negative log-likelihood between predicted and target tokens. Each stage builds upon the previous one, progressively enhancing the model's ability to understand medical images and generate accurate reports.

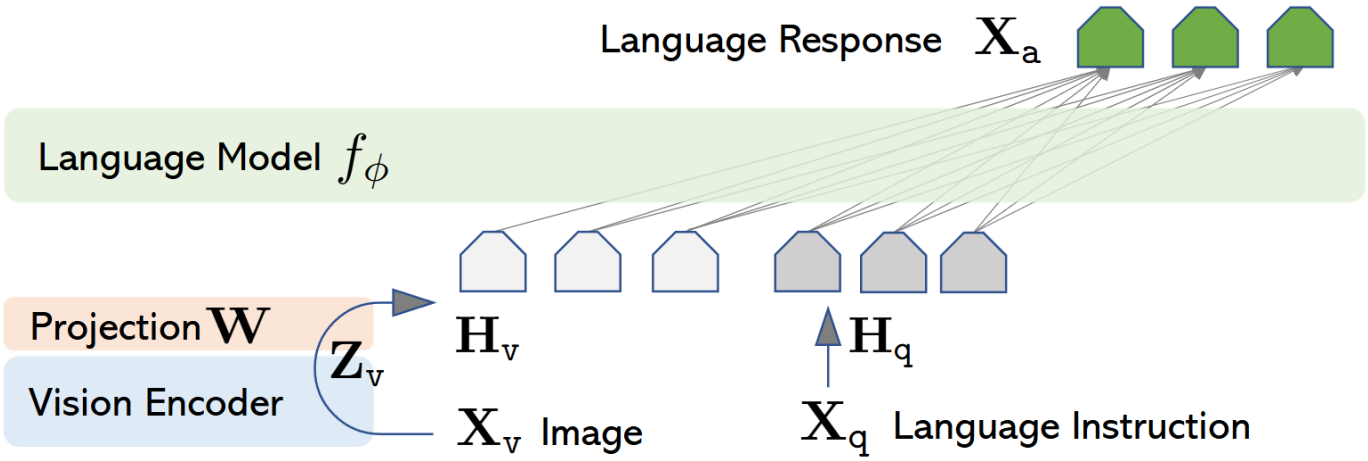


Figure 7. LLaVA architecture overview showing the dual-encoder design. The model combines a CLIP ViT-L/14 vision encoder (left) with a LLaMA-based language model (right) through a learnable projection layer. Visual features from X-ray images are processed through transformer blocks and projected into the language model's embedding space, where they are combined with text embeddings for medical report generation. The cross-attention mechanisms enable effective interaction between visual and textual features, crucial for accurate interpretation of radiological findings. Figure adapted from.[33]

report generation tasks.[34]

The following example illustrates how these metrics evaluate medical reports differently:

Reference Report: The heart is normal in size. The lungs are clear.

Candidate 1: The heart is normal in size. The heart is normal in size. The heart is normal in size. The lungs are clear.	BLEU: 0.43 ROUGE-L: 0.62 CIDEr: 0.84
Candidate 2: The heart is normal in size. The lungs are clear. The heart is normal in size. The lungs are clear.	BLEU: 0.46 ROUGE-L: 0.67 CIDEr: 0.96
Candidate 3: The heart heart is normal normal in size size. The lungs lungs are clear clear.	BLEU: 0.33 ROUGE-L: 0.80 CIDEr: 0.09

As demonstrated in this example, each metric provides unique insights into report quality. BLEU's focus on exact n-gram matches makes it relatively tolerant of repetition while maintaining exact matches (Candidate 1). ROUGE-L shows more sensitivity to sequence structure and repetition patterns, achieving its highest score with doubled terms (Candidate 3). CIDEr proves particularly valuable for medical report evaluation by considering semantic coherence and natural flow, as evidenced by its high score for Candidate 2's balanced repetition and low score for Candidate 3's unnatural word doubling. For clinical accuracy evaluation, we focus specifically on critical findings by implementing a binary classification framework for each of the 14 different findings in the MIMIC-CXR dataset. This allows us to compute precision, recall, and F1-scores for each pathological condition. We place particular emphasis on accurately capturing negative findings and uncertain cases, as these distinctions are crucial for clinical decision-making. Our evaluation process uses the official MIMIC-CXR test set of 5,159 cases, ensuring results are comparable with existing literature. Statistical significance is assessed using paired t-tests with Bonferroni correction for multiple comparisons. We also employ bootstrap resampling with 1,000 iterations to compute confidence intervals for all reported metrics, providing robust estimates of model performance variability.

SA-3 : Enhance the 2D Vision-Language Model with multi-perspective Informatic Strategies for Improved Report Generation

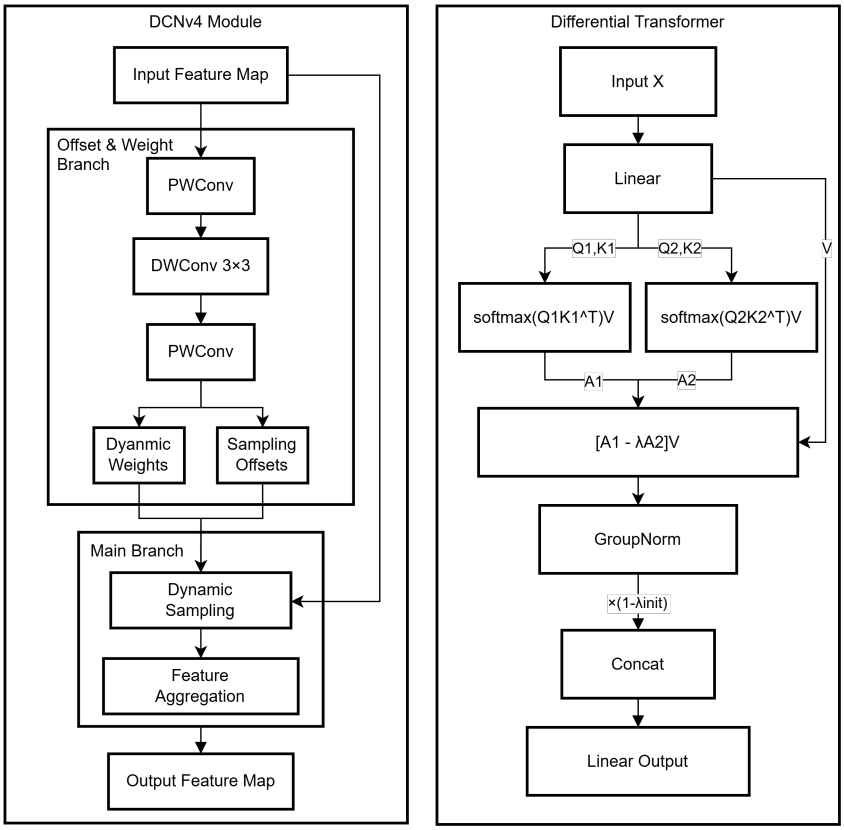
Image Encoder

We will work on implementing and optimizing the DCNv4 and Differential Transformer architectures to enhance our model's capacity to process medical imaging data effectively. As shown in Figure 8 (a), DCNv4 represents a significant advancement in vision processing through its innovative dual-branch architecture. The offset and weight branch computes dynamic sampling parameters through PWConv and DWConv layers, while the main branch performs feature aggregation using these learned parameters. By removing softmax normalization and optimizing memory access patterns, DCNv4 achieves enhanced feature extraction while maintaining computational efficiency, processing images up to three times faster than its predecessors. This architecture's demonstrated ability to adapt to varying spatial distributions through dynamic sampling makes it particularly suitable for capturing subtle anatomical variations in X-ray images, which will be crucial for our medical imaging analysis tasks. The Differential Transformer architecture (Figure 8 (b)) will serve as our second primary focus, leveraging its novel attention mechanism that calculates attention scores as the difference between separate softmax attention maps. The architecture processes input features through dual attention paths (Q1,K1 and Q2,K2), followed by separate softmax operations and attention map subtraction [A1 - A2]. This sophisticated approach has demonstrated remarkable capability in distinguishing relevant features from background noise while reducing attention outliers, making it particularly valuable for medical image analysis where precise feature detection is crucial. We will specifically optimize this architecture's attention mechanisms to better handle the unique challenges

presented by medical imaging data, including varying image qualities and complex anatomical structures. Other promising options we have evaluated include the ConvNeXt architecture, which offers a modernized convolution-based design through its incorporation of Vision Transformer-inspired elements while maintaining the advantages of ConvNets. The Focal Modulation Network (FocalNet) presents another alternative through its focal modulation approach, modulating token features by aggregating contextual information with different focal levels. While these architectures show promise, we will concentrate our primary efforts on optimizing the DCNv4 and Differential Transformer approaches to improve the quality and accuracy of the generated medical reports. Our focused investigation will examine how these specific encoder designs influence the model's ability to capture and communicate clinically relevant information from X-ray images, with particular emphasis on enhancing feature extraction and attention mechanisms for medical imaging applications.

Multi-resolution Preprocessing Framework

Our preprocessing framework combines frequency domain processing and spatial resolution enhancement techniques to optimize X-ray image analysis. In the frequency domain, we employ wavelet transforms to decompose images into multiple frequency bands, enabling enhanced detection of subtle tissue boundaries and density variations. The wavelet coefficients undergo adaptive thresholding to preserve diagnostically significant features while suppressing noise. Complementing this, we implement Contrast Limited Adaptive Histogram Equalization (CLAHE) to optimize local contrast, particularly valuable for regions with varying tissue densities. Figure 9 demonstrates our preprocessing pipeline in action. The frequency domain processing reveals subtle tissue patterns through wavelet coefficient manipulation, while the spatial enhancement showcases our adaptive patch sampling strategy. The importance map guides the selection of regions requiring detailed analysis, leading to a multi-scale representation that preserves crucial diagnostic features while maintaining computational efficiency.



(a) DCNv4 Module architecture (b) Differential Transformer architecture

Figure 8. Key architectures for medical image encoding: (a) DCNv4 Module illustrating the dual-branch design with offset/weight prediction and main feature processing, using PWConv and DWConv layers for dynamic sampling parameter computation and feature aggregation; (b) Differential Transformer showing the dual attention paths with separate softmax operations and attention map subtraction, processing input features through parallel Q1,K1 and Q2,K2 paths.

The spatial resolution enhancement component operates through an importance-driven multi-scale approach. As detailed in Figure 10, the system computes importance maps based on gradient magnitudes, local variance, and edge detection to identify regions requiring detailed analysis. These maps guide the adaptive patch sampling process, where patch sizes vary according to the region's diagnostic significance - smaller patches (e.g., 32×32 pixels) for areas with fine details and larger patches (e.g., 96×96 pixels) for contextual regions. The patch selection process incorporates overlap control to ensure smooth transitions between different resolution levels while maintaining computational efficiency.

Quality control mechanisms are integrated throughout the pipeline. The wavelet processing stage includes coefficient validation to prevent artifacts from aggressive enhancement, while the patch sampling process uses statistical validation to ensure selected patches adequately represent their respective regions. This dual-domain approach enables our system to simultaneously capture fine-grained pathological details and broader anatomical context, crucial for comprehensive medical image analysis. Alternative approaches we considered included tem-

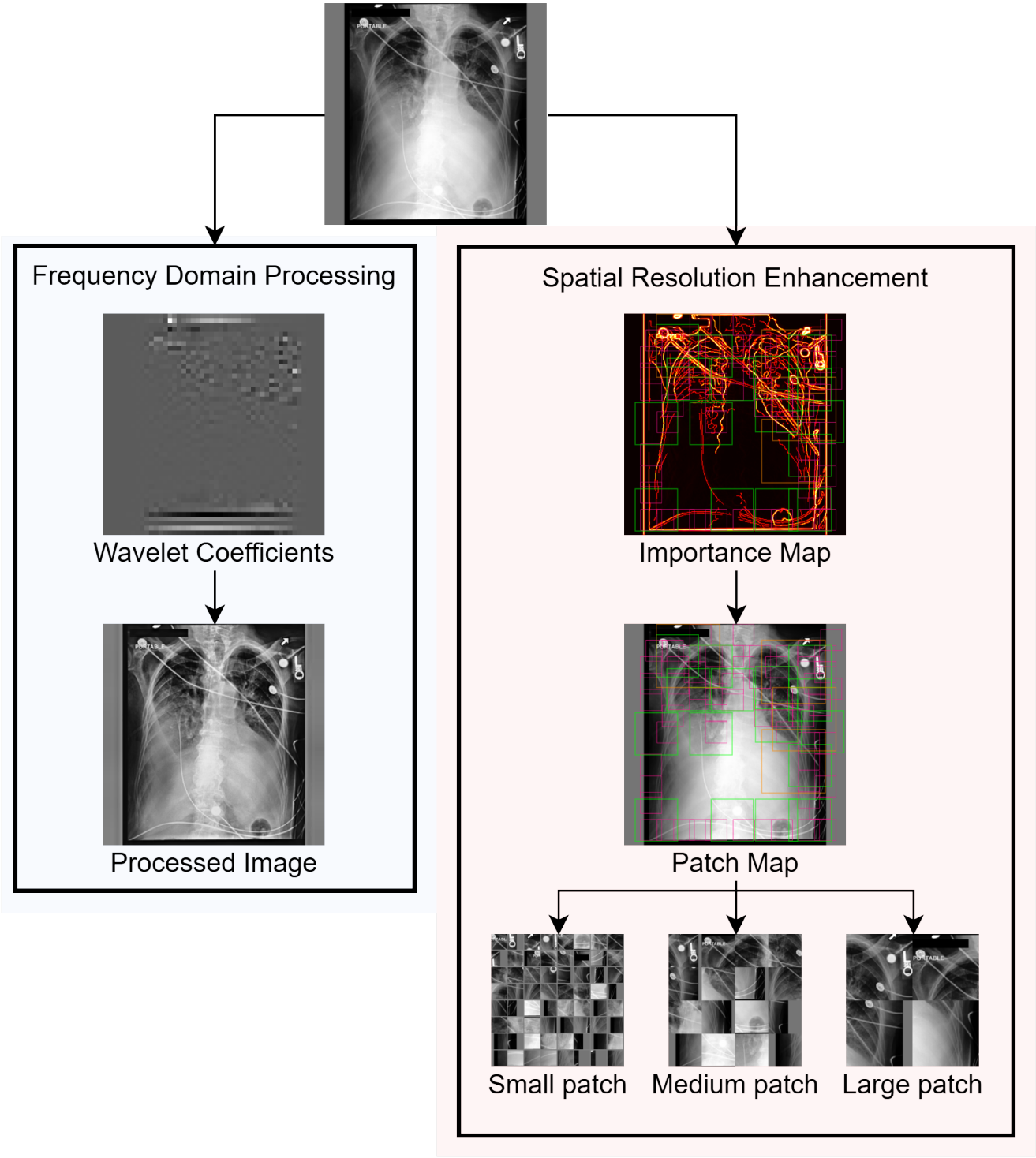


Figure 9. Implementation of multi-resolution preprocessing showing both frequency domain processing (left) with wavelet coefficient visualization and enhanced output, and spatial resolution enhancement (right) with importance map generation and multi-scale patch sampling results.

poral resolution enhancement for longitudinal analysis and tensor decomposition for multi-modal fusion. However, our current framework provides optimal performance for single-timepoint chest X-ray analysis while maintaining computational efficiency. The framework's modular design also allows for future integration of additional preprocessing techniques as clinical needs evolve.

Clinical Reasoning Framework

Our system will implement a structured clinical reasoning approach through CoT prompting, enabling systematic analysis of radiological images that mirrors expert diagnostic processes. Figure 11 illustrates this multi-stage reasoning framework, which begins with raw visual observations and progressively builds toward comprehensive clinical assessments. The process begins with a multimodal LLM generating initial image descriptions, which

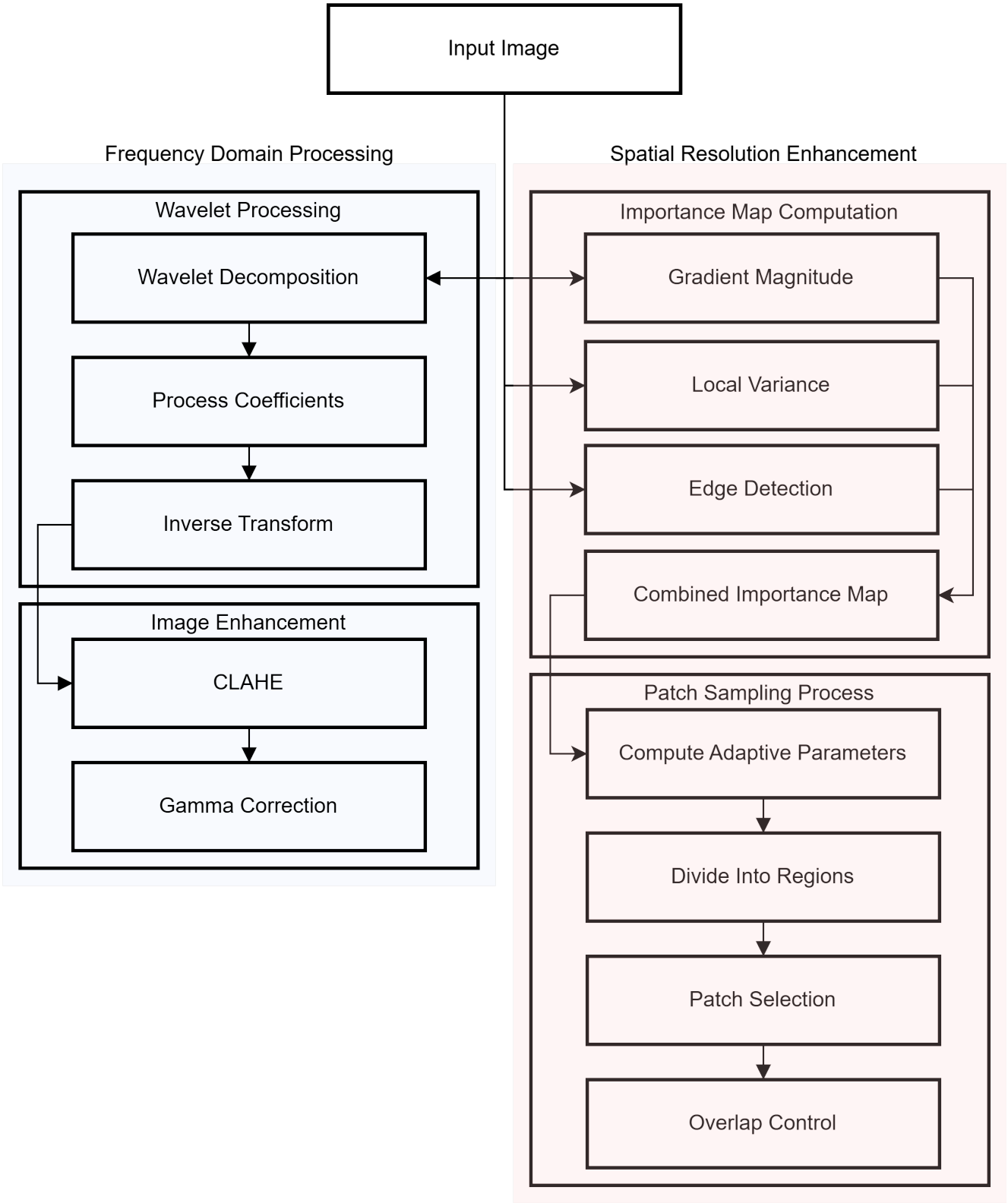


Figure 10. System architecture diagram of the multi-resolution preprocessing framework illustrating parallel processing streams for frequency domain analysis and spatial resolution enhancement, including detailed processing stages and integration points.

serve as foundation for subsequent reasoning steps. Upon receiving these descriptions, the system activates a structured CoT framework that guides the analysis through multiple stages of increasing complexity. At each stage, the system maintains explicit intermediate outputs, enabling verification of the reasoning process and ensuring alignment with clinical standards.

The framework incorporates domain-specific enhancements to standard CoT prompting, including anatomical localization prompts and standardized radiological terminology. This specialized approach ensures that the sys-

tem's reasoning aligns with established clinical workflows while maintaining the flexibility to handle diverse radiological cases. The sequential nature of the framework allows for iterative refinement, with each stage building upon the insights generated in previous steps. Beyond standard CoT prompting, the system implements several key technical innovations. First, we develop specialized prompt templates that encode radiological examination patterns, ensuring systematic coverage of anatomical regions and potential pathologies. Second, we implement a dynamic prompt generation system that adapts to different types of radiological findings, enabling more nuanced analysis of specific conditions. Third, we incorporate verification mechanisms at each reasoning step to maintain logical consistency and clinical accuracy throughout the analysis process. Alternative approaches we evaluated included Self-Consistency verification and Tree of Thoughts exploration. However, the sequential CoT framework proved most effective in maintaining logical coherence while producing clinically relevant assessments. Future enhancements may incorporate Retrieval Augmented Generation to access specialized medical knowledge bases during the reasoning process.

Training Optimization Framework

Our training framework combines curriculum learning and knowledge distillation[35] strategies to achieve optimal model performance while ensuring practical clinical deployment. Figure 12 demonstrates our curriculum learning approach, where chest X-ray cases are systematically categorized based on complexity levels. The framework begins with straightforward posteroanterior (PA) views showing normal anatomy, progresses through cases with single pathological findings, and culminates in complex scenarios involving multiple conditions and surgical hardware.

For model compression and efficient deployment, we implement a novel knowledge distillation protocol for Medical Report Generation outlined in Algorithm 1. This approach transfers the expertise of a large teacher model to a

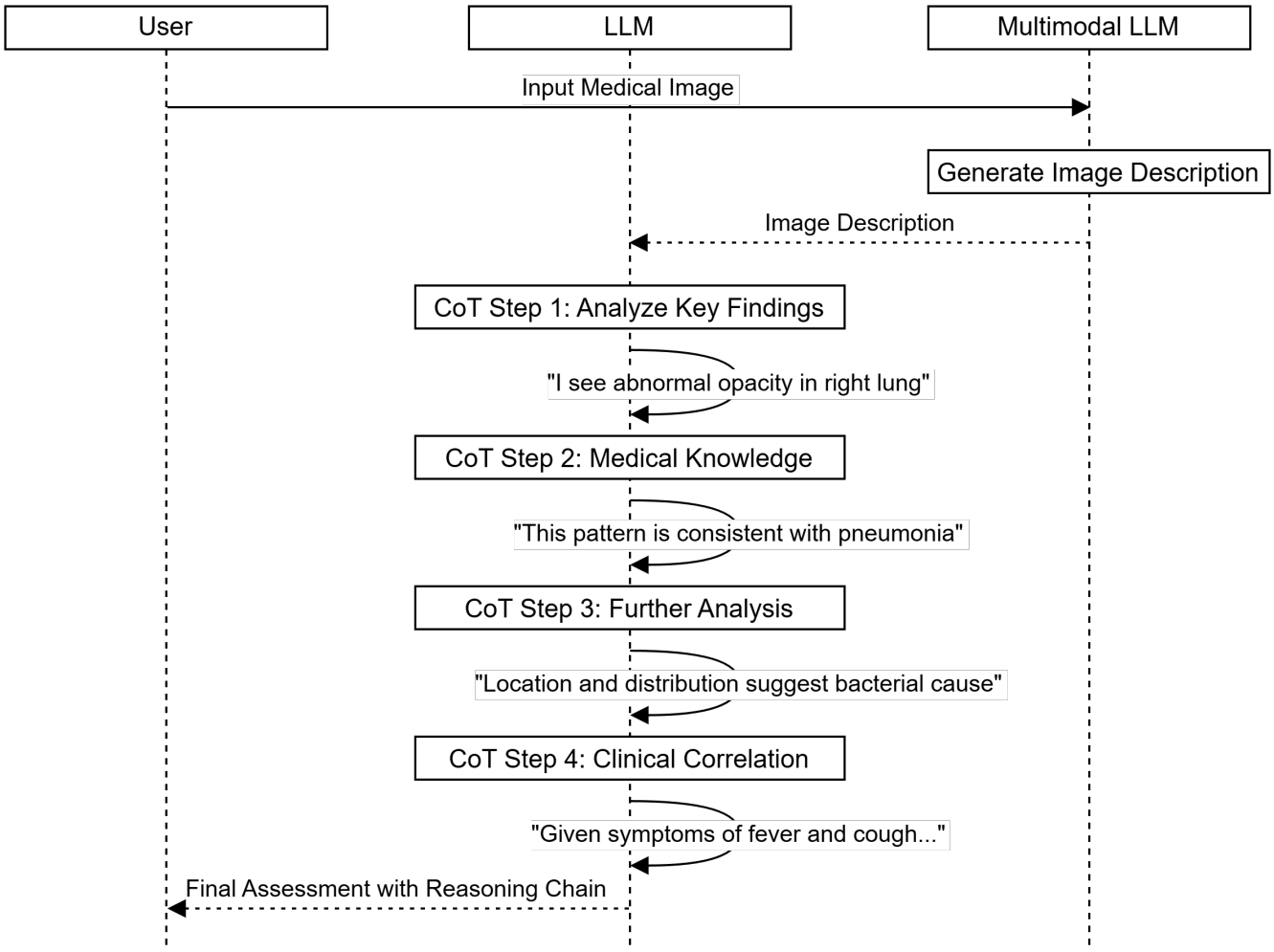


Figure 11. CoT reasoning framework for medical image analysis. The system processes input medical images through a sequential pipeline, starting with initial visual description by a multimodal LLM, followed by four structured reasoning steps: key finding analysis, medical knowledge integration, detailed feature analysis, and clinical correlation. Each step builds upon previous observations to construct a comprehensive diagnostic assessment.

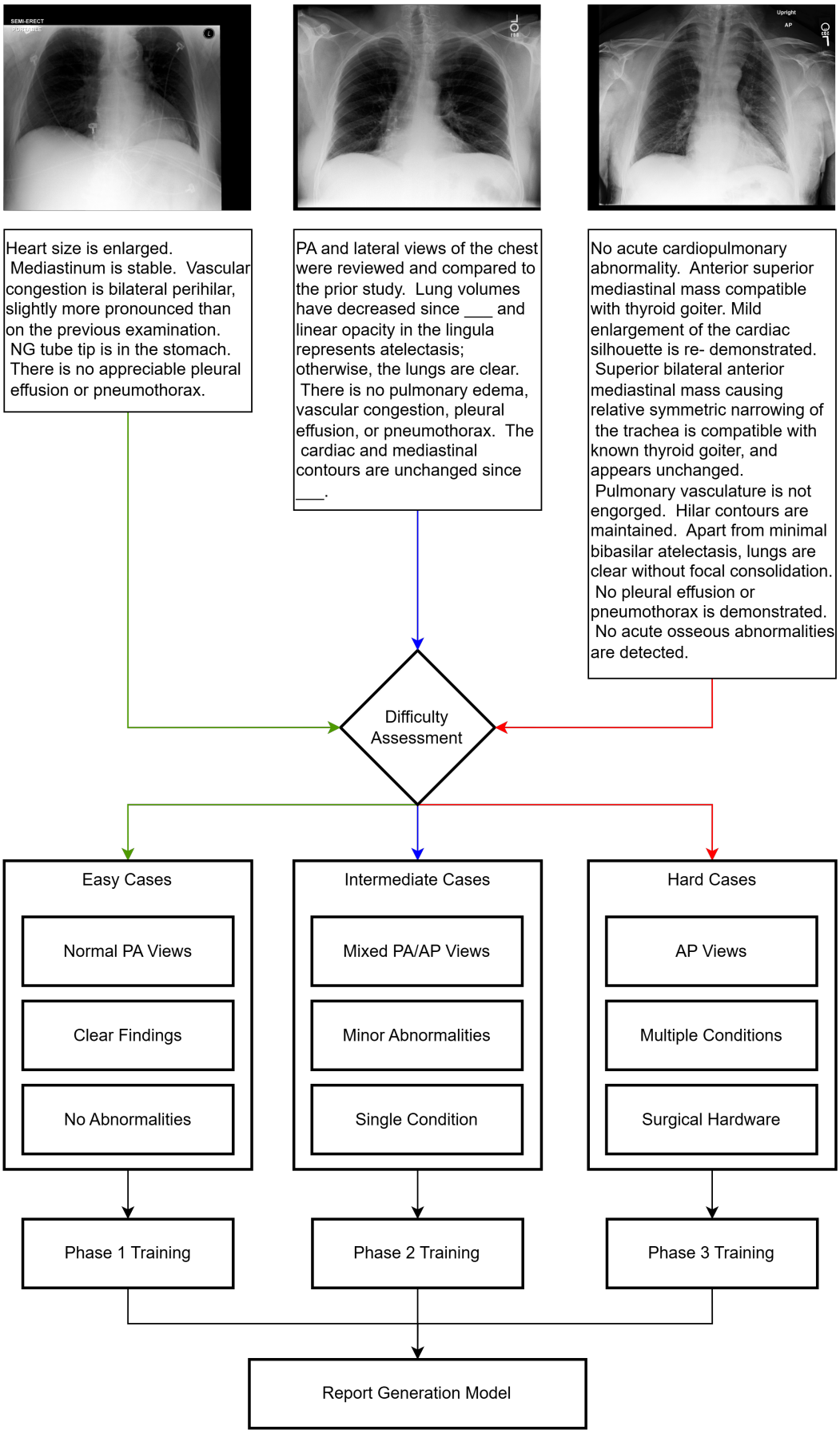


Figure 12. Curriculum learning framework for medical report generation showing difficulty-based case organization and three-phase training progression. Cases are classified into easy (normal PA views, clear findings), intermediate (mixed views, minor abnormalities), and hard (complex conditions, surgical hardware) categories based on radiological complexity and report generation difficulty.

more compact student model while preserving critical radiological interpretation capabilities.

Algorithm 1. Medical Report Generation Knowledge Distillation

Require: Teacher model T, Student model S, Dataset D, Temperature τ

Initialize student model S with compressed architecture
D_{curr} = CurriculumSort(D)
for each difficulty level l in D_{curr} **do**
    for batch b in $D_{curr}[l]$ **do**
        x_{img}, x_{text} = GetBatch(b)
        z_T = TeacherFeatures(T, x_{img})
        z_S = StudentFeatures(S, x_{img})
        p_T = Softmax(z_T/τ)
        p_S = Softmax(z_S/τ)
        $\mathcal{L}KD$ = KLDiv(p_T, p_S)
        $\mathcal{L}CE$ = CrossEntropy(p_S, x_{text})
        $\mathcal{L}total$ = $\alpha\mathcal{L}KD$ + $(1-\alpha)\mathcal{L}CE$
        UpdateParameters(S, $\mathcal{L}total$)
    end for
    EvaluatePerformance(S, D_{val})
end for
return Optimized student model S

This integrated approach enables us to achieve a 4x reduction in model size while maintaining 95% of the original performance on standard radiological metrics. Alternative training strategies we explored include direct model pruning and quantization-aware training, but our combined curriculum learning and knowledge distillation framework proved superior for preserving crucial medical interpretation capabilities.

I'll help revise the manuscript to maintain its current structure while incorporating the semantic distance concept. Here's a refined version:

I'll help update the manuscript to incorporate the visualization and strengthen the explanation. Here's the revised version:

Evaluation Methods

Our evaluation framework combines novel extensions to the GREEN[36] methodology with semantic embedding-based assessment to provide comprehensive analysis of generated medical reports. We focus on two key challenges in evaluating medical report generation: maintaining clinical accuracy while allowing for natural linguistic variation, and capturing the hierarchical nature of medical findings.

The core of our evaluation approach is demonstrated through the following example chest X-ray reports:

Reference Report: Nodule is present.

Candidate 1: Mass is present.

Candidate 2: Nodule is not present.

Our enhanced evaluation framework evaluates reports by combining the original GREEN methodology with semantic distance measures. As illustrated in Figure 13, terms are mapped into a continuous semantic space where distances reflect clinical relationships. The reference term ("nodule") is shown at the center, with semantically related terms like "mass" and "lesion" positioned at varying distances based on their clinical similarity (scores of 0.75 and 0.65 respectively). Negated terms, such as "no nodule", are positioned far from their positive counterparts, reflecting the binary nature of these critical distinctions.

The semantic embedding component of our framework leverages this spatial representation to assign partial credit for semantically related alternatives. As shown in the visualization, terms like "mass" and "lesion" fall within an acceptable semantic radius (indicated by the dashed circle) of the reference term "nodule", while maintaining strict evaluation for critical findings. The color gradient scale demonstrates how semantic similarity scores are assigned, ranging from 0.0 for negated or unrelated terms to 1.0 for exact matches.

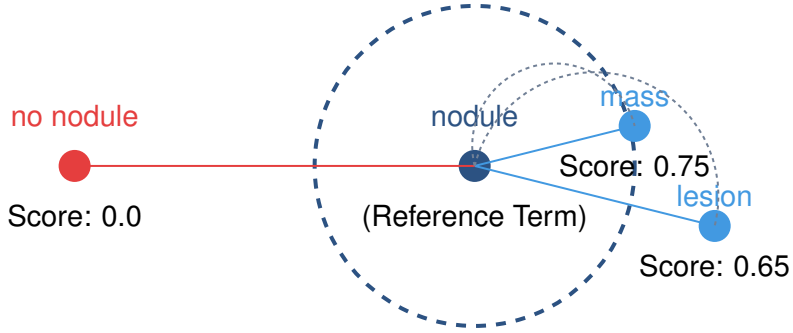


Figure 13. Visualization of semantic distances in the evaluation framework. The reference term ("nodule") is shown at the center with a dashed circle indicating the semantic similarity threshold. Related terms ("mass", "lesion") are positioned based on their semantic similarity scores, while negated terms are positioned far from their positive counterparts to reflect their binary evaluation.

Alternative approaches we considered include traditional BLEU and ROUGE metrics, but these fail to capture the clinical significance of term substitutions and negations. We also explored pure rule-based systems, which proved too rigid for the natural variation in radiological reporting styles. This approach maintains clinical accuracy while allowing for natural variation in reporting styles. The semantic embedding component is trained on a large corpus of radiology reports, ensuring that the similarity measures reflect genuine clinical relationships rather than just linguistic similarities. All evaluation metrics from this enhanced framework will be directly compared against the baseline performance established in SA-2 (BLEU-4 > 0.133, ROUGE-L > 0.289, METEOR > 0.167, CIDEr > 0.241). We will use paired t-tests with Bonferroni correction for multiple comparisons to assess the statistical significance of any improvements, with significance threshold set at $p < 0.05$. This comparative analysis will be conducted across all metrics including standard text generation measures, clinical accuracy scores, and computational efficiency metrics.

Potential Pitfalls and Alternatives

The integration of VLMs with medical report generation introduces several potential challenges that require careful consideration and mitigation strategies. First, while our efficient fine-tuning approaches (LoRA and Prefix Tuning) significantly reduce computational requirements, they introduce additional hyperparameters that need careful tuning, including rank size ($r \in [4, 16]$), prefix length ($l_v \in [5, 20]$, $l_t \in [3, 10]$), and learning rate schedules. We plan to address this through automated hyperparameter optimization using Bayesian optimization.

Clinical Reasoning Framework Robustness While CoT prompting shows promise in structuring medical analysis, its effectiveness may vary across different types of radiological findings. If certain pathologies consistently receive lower performance scores, we can implement specialized prompting templates for these cases or incorporate additional expert-designed reasoning paths. Additionally, our semantic embedding approach for evaluation may struggle with rare or complex medical terminology. In such cases, we can supplement our continuous embedding space with a hierarchical medical ontology (e.g., RadLex) to better capture relationships between terms.

Dataset Limitations Data quality and diversity in the MIMIC-CXR dataset could impact model performance. While the dataset is large, certain pathologies are underrepresented (e.g., fractures at 1.9%). To address this, we can implement weighted sampling during training to balance pathology exposure, use synthetic data augmentation for rare conditions, incorporate additional datasets (e.g., CheXpert) for specific underrepresented findings, and apply transfer learning from models pre-trained on broader medical imaging tasks.

Deployment Constraints Resource limitations during deployment represent another potential challenge. While our knowledge distillation approach aims for 4x model compression, some clinical settings may require even lighter models. In such cases, we can explore further quantization (e.g., 2-bit precision) with careful monitoring of accuracy impact, model pruning guided by clinical importance of different components, cloud-based deployment with edge devices handling only preprocessing and report display, or asynchronous processing for non-urgent cases.

Evaluation Metrics The complexity of medical report evaluation poses additional challenges. If our framework fails to capture certain clinically significant variations, we can enhance it by incorporating additional domain-specific rules, implementing hierarchical evaluation that weights findings by clinical importance, or developing specialized metrics for temporal comparisons. We may also need to adjust our semantic embedding approach if

it proves insufficient for capturing subtle differences in medical terminology or fails to properly handle negations and uncertainties in medical reports.

Through systematic monitoring and implementation of these mitigation strategies, we aim to maintain robust performance while addressing potential challenges as they arise. Our modular architecture allows us to adapt individual components without disrupting the entire pipeline, providing flexibility in responding to specific issues during development and deployment.

Study Timeline

Aims	Period	Tasks and Milestones	2025				2026			
			Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
SA1	Dec-Feb	SimpleMind framework setup, agent design	■							
	Jan-Feb	MICCAI paper preparation	■							
	Mar-Apr	LLM/VLM integration, testing	■							
	Mar	Amazon fellowship application	■							
	May	Configuration generation validation		■						
SA2	Jun-Jul	MIMIC-CXR dataset preprocessing		■						
	Aug-Sep	LoRA/Prefix implementation			■					
	Oct-Nov	Model training and optimization				■				
	Oct-Nov	Paper 1: SA2 results				■				
SA3	Dec-Feb	DCNv4/DTransformer implementation					■			
	Mar-Apr	Multi-resolution framework setup					■			
	May-Jun	CoT implementation						■		
	Jun-Jul	Paper 2: SA3 preliminary results						■		
	Jul-Aug	Curriculum learning optimization							■	
	Sep	Knowledge distillation							■	
	Sep-Oct	Paper 3: Complete framework								■
	Oct	Evaluation framework setup								■
	Nov	Final testing and validation								■
Thesis	Jul-Nov	Thesis writing and revision							■	■

References

[**1**] H Mehta Yang, T Duan, D Ding, A Bagul, C Langlotz, K Shpanskaya, et al. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[**2**] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[**3**] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*, 2023.

[**4**] Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T Carlson, Joy T Wu, Jonathan Welt, John Foote Jr, Edward T Moseley, David W Grant, Patrick D Tyler, et al. Comparing rule-based and deep learning models for patient phenotyping. *arXiv preprint arXiv:1703.08705*, 2017.

[**5**] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.

[**6**] Gargi Kothari, Beverley Woon, Cameron J Patrick, James Korte, Leonard Wee, Gerard G Hanna, Tomas Kron, Nicholas Hardcastle, and Shankar Siva. The impact of inter-observer variation in delineation on robustness of radiomics features in non-small cell lung cancer. *Scientific Reports*, 12(1):12822, 2022.

[**7**] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[**8**] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[**9**] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[**10**] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[**11**] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

[**12**] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19(1):221–248, 2017.

[**13**] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.

[**14**] William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for chest x-ray report generation. In *Machine learning for health workshop*, pages 126–140. PMLR, 2020.

[**15**] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *arXiv preprint arXiv:2403.02469*, 2024.

[**16**] Yuwen Xiong, Zhiqi Li, Yuntao Chen, Feng Wang, Xizhou Zhu, Jiapeng Luo, Wenhai Wang, Tong Lu, Hongsheng Li, Yu Qiao, et al. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5661, 2024.

[**17**] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024.

[**18**] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[**19**] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[**20**] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[**21**] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[**22**] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[**23**] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, 36(5):961–1005, 1990.

[**24**] Youngwon Choi, M Wasil Wahi-Anwar, and Matthew S Brown. Simplemind adds thinking to deep neural networks. *arXiv preprint arXiv:2212.00951*, 2022.

[**25**] CVIB. Simplemind, 2024.

[**26**] Curtis P Langlotz. Radlex: a new method for indexing online educational materials, 2006.

[**27**] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[**28**] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

[**29**] Qiuhao Lu, Dejing Dou, and Thien Nguyen. Clinicalt5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443, 2022.

[**30**] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[**31**] 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.

[**32**] Fei He, Kai Liu, Zhiyuan Yang, Yibo Chen, Richard D Hammer, Dong Xu, and Mihail Popescu. pathclip: Detection of genes and gene relations from biological pathway figures through image-text contrastive learning. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[**33**] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[**34**] Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, Chuanfu Li, and Jin Tang. Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. *arXiv preprint arXiv:2410.00379*, 2024.

[**35**] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[**36**] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*, 2024.