VIETNAM NATIONAL UNIVERSITY HCMC

UNIVERSITY OF INFORMATION TECHNOLOGY



FINAL REPORT

CS338.P22

VISION-LANGUAGE TRANSFORMER FRAMEWORK FOR AUTOMATED MEDICAL IMAGING REPORTING

Faculty: Computer Science Faculty

Instructor: PhD. Duong Viet Hang

|  |  |  |  |
| --- | --- | --- | --- |
| TT | Student Name | StudentID | Email |
|  | Tran Minh Quan | 22521191 | 22521191@gm.uit.edu.vn |

Ho Chi Minh City – 2025, June.

VISION-LANGUAGE TRANSFORMER FRAMEWORK FOR AUTOMATED MEDICAL IMAGING REPORTING

***Abstract:*** *In this study, we propose a unified vision-language transformer framework for automated radiology report generation from chest X-ray images. Our system leverages recent advances in both vision and language modeling, integrating multiple visual encoders-ResNet152, ViT, and CLIP-with a domain-specific language decoder, BioBART. A lightweight multi-layer perceptron (MLP) is used to project visual embeddings into the language embedding space, facilitating seamless cross-modal alignment. We evaluate the framework using the IU-Xray dataset, which is a small-scale but clinically relevant benchmark. Experimental results show that transformer-based models, particularly CLIP paired with BioBART, outperform traditional CNN-RNN approaches in both fluency and semantic relevance. Ablation studies further demonstrate that a shallow MLP architecture yields more stable performance under low-resource conditions. These findings highlight the potential of vision-language transformers for enhancing clinical workflows through automated, accurate, and coherent radiology reporting.*

# Introduction

Medical imaging plays a vital role in the diagnosis, monitoring, and treatment of a wide range of clinical conditions. Chest X-ray, as one of the most commonly used and accessible imaging modalities, is particularly crucial for detecting pulmonary diseases such as pneumonia, tuberculosis, and lung cancer. Despite its ubiquity, interpreting chest X-ray images remains a non-trivial task that requires specialized expertise. Radiologists are trained to identify subtle patterns and abnormalities across different anatomical structures, and the process of report writing can be both time-consuming and prone to human variability. These limitations are exacerbated in settings where radiological expertise is scarce, such as rural or low-resource hospitals, leading to diagnostic delays and inconsistent reporting quality.

To address this bottleneck, the task of automated medical report generation has attracted increasing interest in the field of clinical AI. The goal is to develop models that can generate radiology reports directly from medical images, thereby augmenting clinical workflows, improving diagnostic throughput, and reducing cognitive workload for clinicians. Early approaches to this problem adopted the encoder-decoder paradigm, leveraging convolutional neural networks (CNNs) as visual encoders to extract features from medical images, and recurrent neural networks (RNNs) - especially long short-term memory (LSTM) networks - as language decoders to generate the text sequence [1]. While this class of models showed promising initial results, it was soon discovered that CNNs often lacked the ability to capture long-range spatial dependencies and fine-grained visual context essential for clinical interpretation. Similarly, RNN-based decoders suffered from limitations in learning long-term dependencies in language and tended to produce generic, repetitive descriptions when trained on small datasets.

These challenges sparked a shift toward transformer-based architectures, which have demonstrated remarkable capabilities in both computer vision and natural language processing. Vision Transformers (ViTs), unlike CNNs, model global relationships between image patches using self-attention mechanisms, allowing them to capture fine-grained spatial context and semantic information across the entire image [2]. Moreover, multimodal models like CLIP (Contrastive Language-Image Pretraining) have pushed the boundaries of joint vision-language understanding by aligning visual and textual features in a shared embedding space, making them suitable for tasks that require cross-modal reasoning. On the language side, models like BioBART - a domain-adapted variant of the BART transformer trained on biomedical literature - offer significant advantages in generating coherent, contextually relevant, and medically accurate text. These advances have opened new doors for more powerful and generalizable solutions to the problem of radiology report generation [3].

However, despite these technical innovations, several key challenges remain unresolved. First, medical datasets are typically small in size due to the sensitive nature of healthcare data and the high cost of expert annotation. Unlike general image-captioning datasets such as MS-COCO, which contain hundreds of thousands of image-text pairs, publicly available radiology datasets like IU-Xray contain fewer than 4,000 annotated reports. This low-resource setting creates difficulties for training large models without overfitting. Second, unlike natural image captioning, radiology report generation demands a high degree of factual correctness and domain-specific language. An incorrect or overly general statement in a generated report can lead to severe clinical consequences, such as misdiagnosis or treatment delays. Third, there exists a semantic gap between visual and textual representations: while image encoders output dense, high-dimensional vectors, language decoders require semantically meaningful token embeddings. Aligning these modalities in a medically coherent way is non-trivial, especially when training data is limited.

In this work, we address these challenges by proposing a unified vision-language transformer framework for automated radiology report generation. Our system integrates multiple visual encoders - including ResNet152, ViT-B/16, and CLIP-ViT-B/32 - with a powerful language decoder, BioBART, to generate clinically meaningful text grounded in visual input. To bridge the gap between the two modalities, we introduce a lightweight multi-layer perceptron (MLP) projection module that aligns visual feature vectors with the embedding space of the language decoder. This architecture is trained and evaluated on the IU-Xray dataset, a publicly available benchmark that includes 3,955 radiology reports paired with corresponding chest X-ray images. The dataset's limited size and real-world characteristics (e.g., noisy reports, image variation) make it an ideal testbed for evaluating model performance in low-resource scenarios.

Our experimental results show that transformer-based encoders, particularly CLIP, outperform traditional CNN-based models in capturing global image context and aligning with textual semantics. When paired with BioBART, the resulting system generates fluent and clinically aligned reports, surpassing prior CNN-RNN baselines across evaluation metrics such as BLEU, ROUGE-L, and METEOR. Interestingly, we find that increasing the complexity of the MLP projection module does not necessarily improve performance; in fact, a shallower MLP yields more stable and robust results, especially when data is limited. An ablation study further supports this finding, indicating that architectural simplicity can enhance generalization and convergence stability in multimodal settings.

This study makes the following key contributions. First, we develop a vision-language transformer framework that leverages recent advances in visual and language modeling to generate automated chest X-ray reports. Second, we systematically compare different visual encoders and evaluate their effectiveness in a unified setup, providing insights into the trade-offs between CNNs and transformers. Third, we propose and analyze the impact of a projection module that facilitates modality alignment in a low-resource context. Finally, we conduct both quantitative (via evaluation metrics) and qualitative (via report analysis) experiments to assess the clinical utility and limitations of the generated outputs.

In doing so, this research aims to advance the field of multimodal clinical AI by demonstrating how transformer-based architectures can be effectively harnessed in real-world medical applications. By focusing on a challenging and underexplored problem domain, we hope to provide a foundation for future efforts in building accurate, interpretable, and scalable AI systems for clinical reporting and decision support.

# Related works

In this section, we review the major threads of research that form the foundation for automated medical report generation from radiological images. We divide the literature review into three main parts: (1) vision-language frameworks based on CNN and RNN architectures, (2) transformer-based vision encoders, and (3) biomedical transformer-based language models. Each subsection provides a detailed discussion on key methods, their strengths and limitations, and how our work builds upon or differs from these prior approaches.

## Vision-Language Frameworks with CNN and RNN Architectures

The earliest approaches to medical image captioning and report generation were largely inspired by techniques from natural image captioning. These methods typically employed convolutional neural networks (CNNs) for visual feature extraction and recurrent neural networks (RNNs), such as long short-term memory (LSTM) models, for sequential language generation. The encoder-decoder paradigm became a dominant architectural pattern.

In one of the foundational works, Vinyals (2015) introduced the Show and Tell model, using a CNN-LSTM structure to generate captions for natural images [4]. This design was later adapted for medical images by Jing (2017), who showed that a similar architecture could be trained to generate short descriptive findings for chest X-rays using the IU-Xray dataset [18]. The model consisted of a CNN-based encoder (typically a variant of ResNet or VGG) and a LSTM-based decoder, supervised with cross-entropy loss on a paired image-text corpus.

However, these models faced several limitations when applied to the medical domain. CNNs, while effective at encoding spatial hierarchies, struggled to capture global dependencies across the image, which are crucial for identifying spatially dispersed abnormalities such as interstitial markings or cardiomegaly. Furthermore, LSTMs often failed to maintain coherence over long sequences, especially when tasked with generating full radiology reports as opposed to short captions. This resulted in generic outputs that lacked clinical specificity.

To improve performance, researchers experimented with hierarchical LSTM models (Krause, 2017), attention mechanisms, and coverage models [5]. Xue (2018) proposed a multimodal recurrent model that employed an attention mechanism over both image regions and the textual history to guide generation [6]. Despite modest improvements, these enhancements were still limited by the fundamental sequence-processing limitations of RNNs and the locality bias of CNNs.

## Transformer-Based Vision Encoders

Transformer architectures have revolutionized representation learning in computer vision by moving away from convolutional operations and instead relying on self-attention to model relationships between all parts of the image simultaneously. The Vision Transformer (ViT), introduced by Dosovitskiy (2020), segments an image into fixed-size patches, linearly embeds them, and processes them through standard transformer layers [7]. This design allows ViT to model global dependencies more effectively than CNNs.

Recent works have demonstrated that ViT and similar architectures outperform CNNs in several medical imaging tasks, including classification, segmentation, and anomaly detection. Regmi (2023) showed that ViT models trained on chest X-rays could more accurately identify pneumonia and pleural effusions compared to traditional CNNs [8]. This superior performance is attributed to ViT’s ability to encode complex contextual information without being constrained by kernel size or layer depth.

An extension to ViT in the multimodal domain is the CLIP model (Radford, 2021), which was trained on 400 million image-text pairs using contrastive learning [9]. CLIP learns to align visual and textual representations in a shared embedding space, making it highly suitable for tasks that involve image-to-text or text-to-image translation. Although CLIP was not trained on medical images, its generalization capabilities allow it to be finetuned or adapted via lightweight projection heads for clinical applications.

In the context of radiology report generation, these transformer-based encoders provide a powerful alternative to CNNs. They offer better global reasoning, are more compatible with transformer-based language decoders, and often require fewer architectural hacks (like attention or hierarchical decoding) to achieve fluent outputs. Our proposed framework incorporates both ViT-B/16 and CLIP-ViT-B/32 as vision backbones, and our experiments confirm their superior performance under low-resource conditions.

## Transformer-Based Biomedical Language Models

Natural language processing in the biomedical domain poses unique challenges due to the complexity and specificity of medical terminology. While general-purpose models like GPT-2 [11] or BERT have achieved remarkable success in open-domain tasks, their performance often degrades in clinical settings due to a lack of domain exposure.

To mitigate this issue, domain-adapted language models have been developed. BioBERT, ClinicalBERT, and BlueBERT are variants of BERT pre-trained on biomedical literature and electronic health records. These models improve downstream performance on tasks like named entity recognition, relation extraction, and question answering.

For text generation tasks, the introduction of BioBART marks a significant milestone. BioBART is a biomedical adaptation of Facebook’s BART model, which uses a denoising autoencoder framework for sequence-to-sequence learning [10]. BioBART is pretrained on large biomedical corpora, such as PubMed abstracts and PMC articles, enabling it to generate text that is both fluent and domain-appropriate.

BioBART’s architecture includes a transformer-based encoder-decoder structure with 12 layers each, and supports input sequences up to 1,024 tokens. It is particularly well-suited for radiology report generation, which demands long-range coherence and accurate reproduction of technical language. In our study, BioBART serves as the language decoder, generating textual output based on projected visual features from the MLP module.

In addition to its architectural advantages, BioBART’s use of biomedical-specific tokenization and vocabulary allows it to better understand abbreviations, synonyms, and clinical phrasing commonly found in radiology reports. Compared to general language models, BioBART achieves superior performance in both semantic alignment and lexical fidelity.

Our work builds upon these advances by integrating BioBART into a unified vision-language framework for medical report generation. We show that even without extensive finetuning, BioBART can produce clinically meaningful reports when provided with rich, context-aware visual embeddings. Furthermore, we analyze how different visual encoders influence BioBART’s output quality and investigate the role of MLP-based feature projection in enabling effective cross-modal alignment.

Overall, the evolution from CNN-RNN pipelines to transformer-based vision-language architectures marks a paradigm shift in automated medical report generation. By combining transformer encoders for both image and text modalities, and aligning them through a carefully designed projection module, our framework leverages the latest research to advance the state-of-the-art in clinical AI.

# Data

The foundation of any machine learning-based approach in medical imaging hinges critically on the quality, relevance, and structure of the data used. In our study, we utilized the IU-Xray dataset, one of the most widely adopted public benchmarks for automated medical report generation. This section provides a comprehensive overview of the dataset, details the preprocessing steps, and discusses the unique challenges of working with low-resource clinical data.

The IU-Xray database was developed by the U.S. National Library of Medicine. It contains chest X-ray images paired with corresponding diagnostic reports written by radiologists. These reports follow a relatively structured format, typically consisting of multiple sections such as “Findings”, “Impression”.The dataset includes 7,470 images and 3,955 unique reports, where each report corresponds to one or two X-ray images from a single patient exam: a frontal (posterior-anterior or PA) view and a lateral view. These dual-view images allow richer visual context and are consistent with standard radiological practice.

However, the dataset presents several inherent limitations that must be carefully addressed before model training. First, the reports were not curated specifically for machine learning, leading to significant variability in writing style, terminology, and structure. Second, the images vary in resolution, contrast, and noise level, reflecting real-world clinical conditions. Third, the dataset is relatively small by deep learning standards, especially when compared to general vision-language datasets like MIMIC-CXR, MS-COCO or Conceptual Captions. For transformer-based models that typically require hundreds of thousands or millions of samples, the IU-Xray dataset poses a clear low-resource challenge.

To prepare the dataset for training, we applied a multi-stage data cleaning and preprocessing pipeline aimed at ensuring consistency, maximizing signal-to-noise ratio, and enabling effective model supervision. The first step involved removing incomplete entries, specifically those reports that were missing the “Findings” section, which serves as the primary target output in our system. This filtering step reduced the total number of usable images from 7,470 to 6,473.

Next, we retained only those patient cases where both frontal and lateral views were available. This constraint ensures uniformity in the input format and avoids bias during training, where the model might learn from view-specific features that do not generalize across different image combinations. After applying this filter, the dataset was further reduced to 5,500 image-report pairs.

To facilitate robust evaluation, we split the cleaned dataset into three subsets: training (70%), validation (15%), and test (15%). Concretely, this resulted in:

* 2,750 training reports (with 5,500 associated images)
* 563 validation reports
* 394 test reports

This split was performed at the patient level to prevent data leakage, ensuring that images from the same patient do not appear in both training and test sets. The relatively small size of the test set reflects the dataset’s limited scale but remains sufficient for comparative evaluation across different model configurations. For the image preprocessing pipeline, we followed best practices in medical computer vision. All X-ray images were resized to 224×224 pixels, a standard input size for pretrained vision backbones like ResNet152 and ViT-B/16.

 On the textual side, we extracted only the “Findings” section of each report, as this portion contains the most diagnostically rich content. The “Impression” section, though valuable, was excluded to maintain uniformity and because many reports lack a clearly defined impression paragraph. Reports were lowercased, punctuation was preserved, and tokenization was performed using the BioBART tokenizer, which supports biomedical terminology and clinical shorthand. The maximum token length was capped at 128 tokens based on distributional analysis of report lengths; over 95% of reports in the dataset fall below this threshold.

A unique characteristic of the IU-Xray dataset is the imbalance in content granularity across different reports. Some entries describe multiple pathologies with high specificity, while others provide general observations. This variability introduces label entropy, where a visually similar image may map to multiple valid textual descriptions depending on the radiologist's interpretation. Addressing this challenge requires models that are capable of generating fluent, semantically accurate, and context-sensitive narratives, rather than simply mimicking surface-level lexical patterns.

Another issue stems from the distribution of diagnostic labels. Although the dataset does not include explicit labels, prior studies have mapped common findings to pathology categories such as cardiomegaly, pleural effusion, or consolidation. The frequency distribution is highly skewed, with “normal” findings being vastly overrepresented compared to rarer conditions like pneumothorax or mass lesions. This long-tail distribution makes it difficult for the model to generalize to rare or complex pathologies, especially under limited supervision. To mitigate this, we did not apply label balancing or resampling techniques, as these could interfere with the free-form nature of report generation. Instead, we rely on pretrained language models (BioBART) to inject domain knowledge during generation.

From a modality perspective, the dataset enables research into multi-view integration, a relatively underexplored direction in radiology report generation. By combining both frontal and lateral views, the model is encouraged to synthesize information across different perspectives, akin to how human radiologists conduct their assessments. In our architecture, visual features from both views are extracted using a shared encoder and then concatenated before being passed through the MLP projection layer. This ensures that both views contribute jointly to the report, potentially improving coverage and reducing false negatives.

The dataset also presents interesting considerations for temporal and structural consistency. Unlike captioning datasets where each image is associated with a single sentence, radiology reports often follow a narrative structure-first describing the anatomy, then observations, followed by clinical conclusions. This sequential dependency is difficult to capture with models trained on flat, non-hierarchical representations. While our current approach treats report generation as a flat sequence-to-sequence problem, future work could explore hierarchical modeling strategies that better mirror the underlying discourse structure of radiology reports.

In summary, the IU-Xray dataset provides a valuable but challenging testbed for vision-language models in clinical settings. Its small size, noisy annotations, and rich semantic content simulate the kinds of constraints faced in real-world applications, making it a compelling choice for benchmarking multimodal learning systems. Our data preprocessing pipeline ensures that the inputs are standardized, the training set is balanced across views, and the model is trained in a way that emphasizes generalization and robustness over memorization. The careful design of this data pipeline is essential for enabling the transformer-based framework to effectively bridge the visual and linguistic domains in a clinically meaningful way.

# Methods

In this section, we present our proposed framework for automated medical report generation from chest X-ray images. The design of this system is motivated by the need to bridge the semantic gap between complex visual input and structured clinical text output, particularly under the constraints of low-resource data settings common in healthcare applications. To this end, we adopt a modular vision-language transformer architecture that combines pretrained vision encoders with a domain-specific language decoder via a lightweight cross-modal projection mechanism.

At a high level, the goal of the system is to generate a coherent, clinically accurate radiology report given a pair of chest X-ray images: one frontal (posterior-anterior) and one lateral view. These images are first processed by a visual encoder to extract rich feature representations that capture both anatomical structures and potential pathological patterns. The extracted features from the two views are then fused and passed through a multi-layer perceptron (MLP) projection module, which aligns the visual features with the embedding space of the language model [11]. Finally, a transformer-based language decoder, pretrained on biomedical text, generates a diagnostic report conditioned on the projected visual features.

Our architectural design is guided by several core principles:

* *Leverage of pretrained knowledge:* We utilize powerful pretrained encoders such as ResNet152, Vision Transformer (ViT-B/16), and CLIP-ViT-B/32 to obtain robust visual features. These models are chosen for their proven ability to generalize well, even in out-of-domain or small-data scenarios. On the language side, we adopt BioBART-a biomedical adaptation of the BART sequence-to-sequence transformer model due to its strong performance in clinical text generation and understanding.
* *Separation of modality-specific components:* To enable flexible experimentation and modular improvement, the system is designed such that the vision and language components are loosely coupled via a projection module. This not only simplifies training but also allows for independent tuning of each part.
* *Minimal training footprint:* In order to prevent overfitting and ensure reproducibility in low-resource settings, only the projection module and the language decoder are fine-tuned during training. The visual encoder is frozen, acting solely as a feature extractor.
* *Multi-view fusion:* A clinically inspired design decision is to process both frontal and lateral chest X-ray images together, reflecting how radiologists routinely interpret imaging studies. Rather than treating each view independently, we encode them jointly to provide the model with a more comprehensive perspective of thoracic structures.

The overall pipeline is depicted in Figure 1, illustrating the three key components of the system:

* *Visual Encoding:* Both the frontal and lateral X-ray images are passed through a shared visual encoder. Depending on the experimental setting, this encoder can be ResNet152 (CNN-based), ViT (pure transformer), or CLIP (multimodal vision-language pretrained model). The output consists of two high-dimensional vectors representing each view.
* *Projection* *Module (MLP):* The two vectors are concatenated into a single joint representation and then passed through an MLP that performs nonlinear transformation and dimensional alignment. This module serves to project the visual features into the same space as the BioBART decoder’s input embeddings.
* *Language Decoding (BioBART):* The output of the projection module is used to initialize the decoder. BioBART then generates the report in an autoregressive manner, predicting each word token based on the previously generated ones and the conditioning image-derived input.

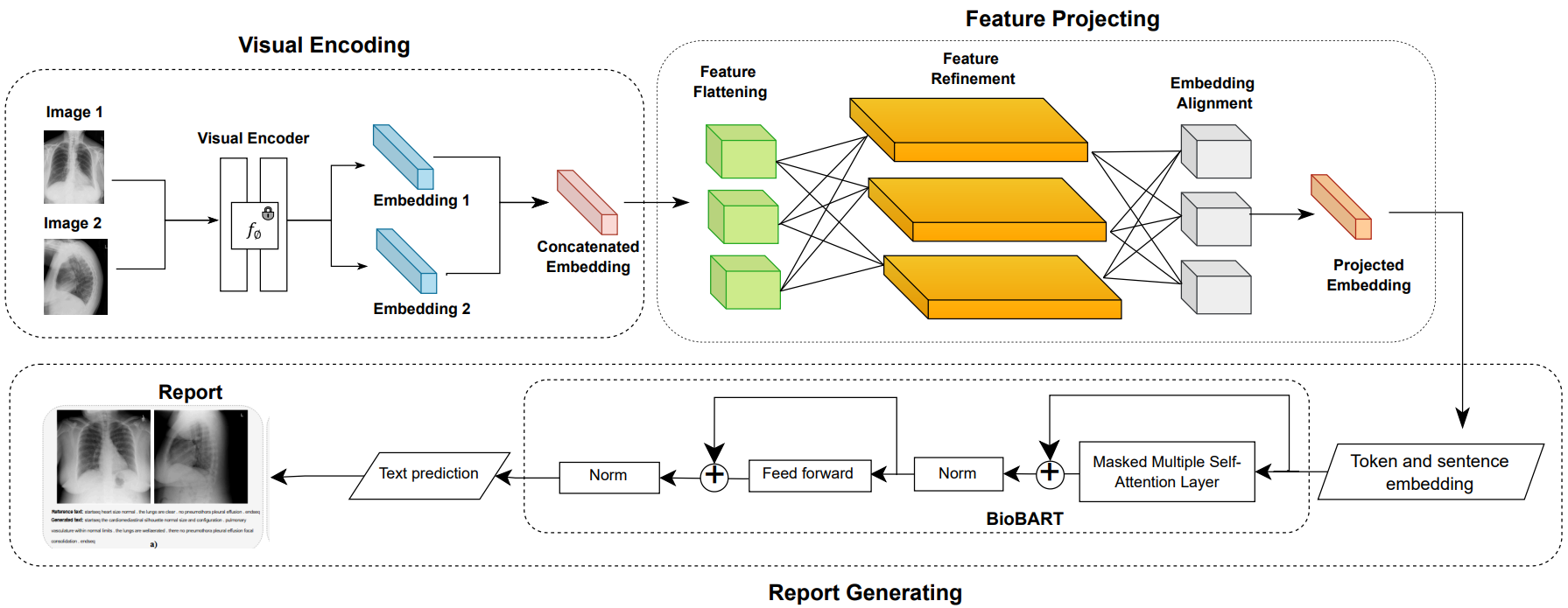


Fig 1. Architecture of the proposed pipeline

This architecture provides several advantages over traditional encoder-decoder approaches that use CNN-RNN combinations. By leveraging transformers in both modalities, the system can capture long-range dependencies and contextual cues more effectively. The use of a shallow MLP as a bridge ensures both model simplicity and computational efficiency, which are important for deployment in clinical settings where resources may be limited.

Moreover, this framework allows for easy ablation and benchmarking. By swapping different visual encoders while keeping the rest of the architecture fixed, we can systematically evaluate their impact on final report quality. Similarly, the projection module can be modified in terms of depth or activation functions to study its influence on alignment performance. In this way, the method not only serves as a practical tool for automated reporting but also as a research framework for probing multimodal learning under data constraints.

The remainder of this section is structured as follows. In Section 5.1, we describe the visual encoding component in detail, comparing the different architectures employed. In Section 5.2, we present the design and implementation of the MLP projection module. Section 5.3 focuses on the report generation process using BioBART, including decoding strategies and training details. These subsections provide a comprehensive view of our methodology for bridging chest X-ray interpretation with natural language generation.

## Visual Encoding

In the task of medical report generation from chest X-rays, the visual encoder serves as the first and one of the most critical components in the pipeline. Its primary role is to extract informative, high-level representations from raw radiographic images that capture the underlying anatomical structures and pathological cues necessary for generating accurate clinical narratives. These representations must be semantically rich and spatially contextual in order to support meaningful cross-modal alignment with the textual domain. Given the medical nature of the task, the encoder must also be robust to the variability and noise common in clinical imaging, such as differences in patient positioning, lighting artifacts, and subtle diagnostic indicators.

|  |  |
| --- | --- |
|  | (1) |

To accommodate these requirements, our framework supports three types of visual encoders, each representing a different family of deep learning architectures: (1) Convolutional Neural Networks (CNNs) via ResNet152, (2) pure transformer-based models via ViT-B/16, and (3) vision-language pretrained models via CLIP-ViT-B/32. All encoders operate over a dual-view input, comprising both frontal and lateral chest X-ray images, which are independently processed through the same encoder architecture. The resulting features are concatenated before being passed to the projection module, allowing the system to integrate information across views.

|  |  |
| --- | --- |
|  | (2) |

### Resnet152

Among the visual encoders considered in our proposed framework, ResNet152 serves as the representative architecture from the convolutional neural network (CNN) family. ResNet152 has gained wide recognition for its powerful hierarchical feature extraction capabilities, generalization performance across domains, and architectural robustness, particularly in low-resource training regimes [13]. These properties make it a strong candidate for medical imaging tasks, such as chest X-ray interpretation, where visual patterns are subtle and datasets are often limited in size.

ResNet152 is a deep residual network comprising 152 layers. Its primary architectural innovation lies in the use of residual connections, which allow the model to learn residual functions instead of direct mappings. These identity shortcuts help prevent the vanishing gradient problem and enable the training of much deeper networks than traditional CNNs. As illustrated in Figure 2, a typical residual block (also referred to as a bottleneck block) consists of three convolutional layers: a 1×1 convolution that reduces dimensionality, a 3×3 convolution for spatial processing, and another 1×1 convolution to restore dimensionality. The input is added to the output of this sequence via a shortcut connection, allowing gradients to propagate directly across layers.

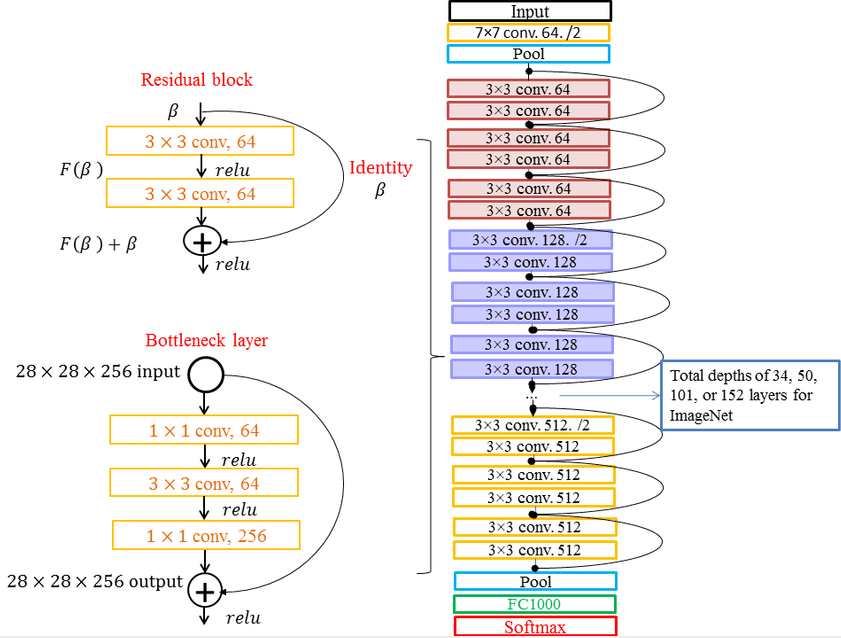


Fig 2. The architecture of ResNet152 with residual bottleneck blocks.

In the full architecture of ResNet152, the input image is first passed through a 7×7 convolutional layer with stride 2, followed by batch normalization and a ReLU activation function. A 3×3 max pooling layer further reduces the spatial dimensions. This is followed by four sequential stages (Conv2\_x through Conv5\_x), each containing multiple bottleneck residual blocks. These stages extract progressively deeper and more abstract representations of the input. At the end of the network, a global average pooling layer is applied to aggregate spatial information into a single 2048-dimensional vector. This vector serves as the visual representation used in downstream tasks.

In our report generation pipeline, each patient case includes two chest X-ray images: a frontal view (posterior-anterior) and a lateral view. Both images are independently passed through the same ResNet152 encoder. Each image is resized to 224×224 pixels and normalized to match the input requirements of the pretrained model. The output of the global average pooling layer is a 2048-dimensional feature vector representing each view. These two vectors - one for the frontal view and one for the lateral view are then concatenated to form a joint representation of 4096 dimensions. This combined vector is subsequently passed through a projection module (described in Section 5.2) to be aligned with the input space of the language decoder.

To ensure training stability and avoid overfitting on the small IU-Xray dataset, the ResNet152 encoder is frozen during training. That is, its weights are kept fixed, and only the downstream MLP projection module and the language decoder are optimized. This strategy leverages the general visual knowledge captured by ResNet152 during its pretraining on ImageNet, enabling transfer learning to the medical domain. Despite the domain difference between natural and medical images, early convolutional layers of pretrained CNNs tend to learn low-level edge detectors and texture patterns that are transferrable to radiological images. This transfer is especially beneficial in low-data regimes, where training an entire encoder from scratch would be computationally expensive and likely ineffective.

ResNet152 is well-suited to the characteristics of chest X-ray data. It captures both local and mid-level features effectively, which are useful for identifying anatomical structures such as the lung fields, diaphragm, ribcage, and heart silhouette. The hierarchical nature of its convolutional layers allows it to detect complex spatial relationships and anatomical abnormalities such as consolidations, pleural effusions, or enlarged cardiac contours. Moreover, the fixed receptive field and weight-sharing properties of CNNs make ResNet152 robust to local distortions and minor variations in imaging protocols, which are common in real-world clinical datasets.

However, ResNet152 also has limitations when applied to tasks that require global spatial reasoning. Because convolutional operations are inherently local, the network may struggle to capture long-range dependencies and interactions between spatially distant regions. This becomes a challenge in detecting bilateral patterns, subtle asymmetries, or diffuse abnormalities, all of which are clinically relevant in chest radiology. Additionally, the lack of a native mechanism for modeling relationships between patches limits its ability to reason over global context - something that transformer-based models such as ViT and CLIP are designed to handle.

Despite these drawbacks, ResNet152 remains a highly effective encoder in the context of our framework. It is computationally efficient, well-understood, and highly compatible with existing infrastructure. In our experiments, the ResNet152-based pipeline produced competitive results across multiple evaluation metrics, including BLEU, ROUGE, and METEOR. Its performance establishes a solid baseline for comparison against more complex models, helping to isolate the contribution of architectural advancements introduced by vision transformers or multimodal pretraining.

In summary, ResNet152 serves as a reliable and interpretable visual encoder for automated chest X-ray report generation. Its residual architecture allows for deep feature extraction, while its pretrained weights enable effective knowledge transfer to the medical imaging domain. Although it lacks the capacity for global attention, its local processing strength and simplicity make it a valuable component in scenarios where training resources are limited or where interpretability and efficiency are paramount. In our unified framework, it forms the basis for understanding how traditional CNNs perform in comparison to transformer-based alternatives when paired with domain-specific language models such as BioBART.

### ViT-B/16

The Vision Transformer (ViT) represents a significant paradigm shift in visual representation learning. Unlike conventional convolutional neural networks (CNNs), which rely on hierarchical filters and local receptive fields, ViT models an image as a sequence of patches and processes these patches using self-attention mechanisms similar to those used in natural language processing [9]. This architectural innovation enables the model to capture long-range dependencies and contextual relationships across the entire image, making it especially suited for tasks that require a global understanding of visual data such as the interpretation of chest X-rays in a clinical setting.

In this work, we adopt the ViT-B/16 variant, which refers to the base-sized Vision Transformer with a patch size of 16×16 pixels. The overall structure of the model is conceptually simple yet remarkably powerful. Given an input image of size 224×224×3, ViT first divides the image into non-overlapping patches of size 16×16. This results in a total of 14×14 = 196 patches. Each patch is flattened into a 1D vector and projected into a 768-dimensional embedding space via a linear transformation. A learnable [CLS] token is prepended to this sequence, which acts as an aggregate representation of the entire image. The resulting sequence of 197 tokens (1 [CLS] + 196 patches) is then passed through a series of transformer encoder blocks, each consisting of multi-head self-attention layers, feed-forward networks, residual connections, and layer normalization. The final output corresponding to the [CLS] token is used as the global image representation for downstream tasks.

ViT’s reliance on self-attention mechanisms rather than convolutional operations allows it to directly model relationships between any pair of image patches, regardless of their spatial proximity. This is particularly beneficial in medical imaging, where pathological findings may appear in spatially distant but semantically correlated regions (e.g., bilateral opacities, asymmetrical lung volumes, or mediastinal shifts). Moreover, the uniform architectural design of transformers simplifies integration with language models, which are typically built on the same self-attention principles. This architectural homogeneity supports more natural cross-modal alignment, especially when paired with a transformer-based decoder such as BioBART.

In our pipeline, ViT-B/16 serves as the visual encoder for both frontal and lateral chest X-ray images. Each image is preprocessed by resizing it to 224×224 pixels and normalizing its pixel values to the [0, 1] range. These preprocessed images are then independently passed through the ViT encoder, which produces a 768-dimensional vector from the final [CLS] token. This vector is considered a compact, context-aware summary of the input image. For each patient case, we obtain two such vectors: vfv\_fvf​ for the frontal view and vlv\_lvl​ for the lateral view. These two vectors are concatenated to form a 1536-dimensional joint embedding.

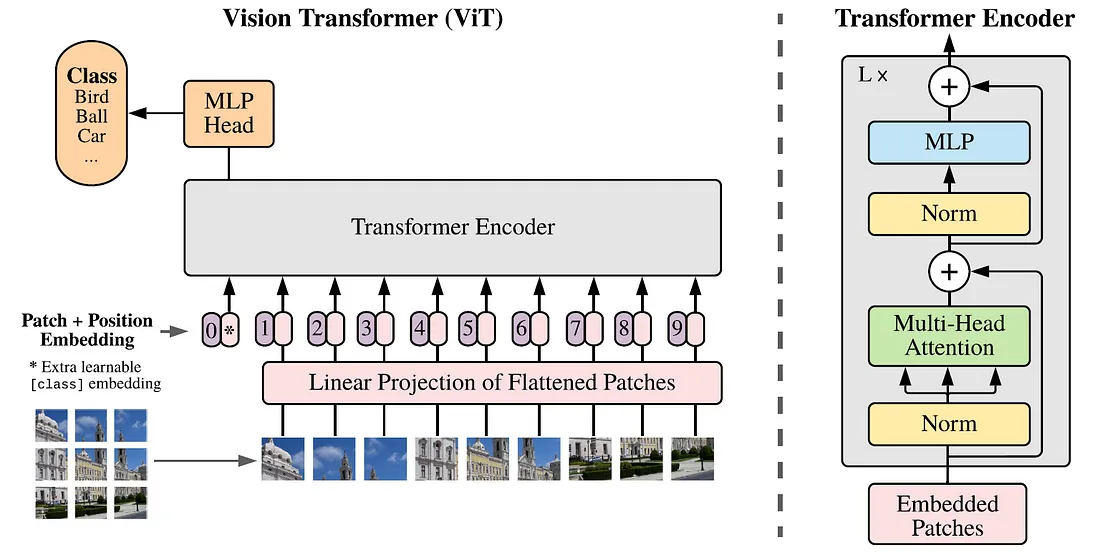


Fig 3. Vision Transformer (ViT) Netwwork Architecture

The joint embedding is then passed to the projection module, which maps it into the input space of the language decoder. As with the ResNet152 encoder, we freeze the weights of the ViT-B/16 encoder during training. This approach leverages the pretrained knowledge acquired on large-scale image datasets (ImageNet-21k or JFT-300M) while avoiding overfitting due to the limited size of the IU-Xray dataset.

One of the defining advantages of ViT in the medical domain is its ability to model global context with minimal inductive bias. Unlike CNNs, which assume that spatially nearby pixels are more related, ViT allows every patch to attend to every other patch through multi-head self-attention. This mechanism enables the model to detect complex spatial patterns that span the entire thoracic cavity, such as diffuse interstitial markings, airspace opacities, or structural shifts due to cardiomegaly. This capability is especially important in chest radiography, where abnormalities often affect bilateral or contralateral structures and where holistic reasoning is essential for accurate diagnosis.

However, the strengths of ViT come with certain trade-offs. Transformers generally lack some of the strong inductive biases that make CNNs effective in low-data regimes, such as locality, translation invariance, and multi-scale structure. While these biases can be useful in focusing the model on relevant image features, their absence in ViT means that the model must learn them from data. This makes training ViT from scratch data-intensive. To mitigate this, we utilize pretrained ViT-B/16 weights, which were learned on large natural image corpora. Even though these pretraining datasets do not contain medical images, previous research has shown that features learned by ViT generalize well to medical tasks when fine-tuned or used in a frozen setting with adapted downstream modules.

In practice, ViT’s high expressivity enables the system to generate more semantically aligned and context-aware reports. Our experiments show that when paired with the BioBART decoder, ViT-based encodings lead to outputs that demonstrate improved coherence and lexical diversity compared to CNN-based encodings. For instance, the model is better able to distinguish between similar clinical findings and more accurately reflect subtle differences in image content.

The use of ViT also supports the design of more interpretable attention maps. Since the self-attention weights in ViT are explicitly computed between all image patches, these can be visualized to understand which regions the model attends to when generating specific tokens in the report. Such visualizations not only provide insights into model behavior but also enhance trustworthiness in clinical applications by allowing practitioners to verify that model predictions are grounded in relevant image regions.

Our implementation of ViT-B/16 is tightly integrated with the report generation framework. During inference, the model processes both images simultaneously, producing a joint visual embedding that is directly responsible for initializing the decoder’s generation process. Unlike feature map-level integration methods, which require spatial alignment and attention over 2D grids, our approach treats the ViT encoder as a global image summarizer, simplifying integration with text generation modules while maintaining high expressiveness.

Despite its strong performance, ViT-B/16 is computationally more demanding than CNNs like ResNet152. The quadratic complexity of self-attention with respect to the number of patches results in higher memory usage, particularly during inference. However, given the relatively small input size (224×224) and the limited sequence length (197 tokens), this cost remains manageable in our setup. Additionally, the frozen nature of the encoder during training reduces training time and hardware requirements.

To summarize, ViT-B/16 brings a powerful alternative to traditional CNN-based visual encoders in the context of chest X-ray interpretation. Its transformer-based architecture enables the modeling of long-range spatial dependencies and complex global interactions, both of which are critical for generating high-quality medical reports. By leveraging a pretrained ViT model and coupling it with a lightweight projection module and a domain-specific decoder, our system is able to produce coherent and clinically relevant textual descriptions with greater fidelity than previous CNN-RNN approaches. The ViT encoder thus plays a pivotal role in bridging the gap between raw visual input and high-level diagnostic language, advancing the capabilities of automated radiology reporting in low-resource environments.

### CLIP-ViT-B/32

The CLIP-ViT-B/32 encoder represents a new generation of visual representation models that are trained not only to understand images but also to align them meaningfully with natural language. CLIP, which stands for Contrastive Language - Image Pretraining, is a vision-language model developed by OpenAI that has shown remarkable generalization capabilities across a variety of downstream tasks without requiring task-specific fine-tuning [12]. In our framework, CLIP-ViT-B/32 serves as a visual encoder that directly benefits from its pretraining on large-scale image-text pairs, thereby offering rich, semantically aligned visual embeddings that are particularly valuable for report generation tasks involving free-form clinical language.

CLIP operates on a dual-tower architecture, where one tower is a vision encoder (in this case, a Vision Transformer) and the other is a text encoder. These two towers are trained jointly using a contrastive loss to bring semantically related image-text pairs closer in a shared embedding space, while pushing unrelated pairs further apart. The training data for CLIP consists of hundreds of millions of image-text pairs scraped from the internet, encompassing a wide range of natural concepts and visual-linguistic associations.Although medical images are underrepresented in this corpus, CLIP still learns highly transferable representations that generalize well to out-of-distribution domains including radiology due to the diversity and scale of its pretraining data.

The CLIP-ViT-B/32 variant we use is based on a Vision Transformer backbone with a patch size of 32×32 pixels. For an input image of size 224×224×3, this results in a total of 49 non-overlapping patches. Each patch is linearly projected into a 768-dimensional embedding space, and a learnable [CLS] token is prepended to the sequence. As in standard ViT architectures, the sequence is processed through multiple layers of self-attention, layer normalization, and feedforward networks. The final representation of the [CLS] token is then used as the global image embedding. Crucially, this representation is trained to be semantically aligned with natural language descriptions, an advantage that traditional vision-only models like ResNet or even pure ViT do not possess inherently.

In our pipeline, each chest X-ray frontal and lateral image is independently passed through the CLIP visual encoder. The output for each view is a 512-dimensional vector, which is the default dimensionality of CLIP's shared embedding space. These two vectors corresponding to the frontal and lateral views respectively, are concatenated to form a 1024-dimensional joint visual embedding.

This vector is then passed to the projection module, which maps it into the decoder input space, allowing the BioBART model to generate a clinical report conditioned on this semantically enriched visual signal.

One of the major advantages of CLIP-ViT-B/32 in the context of radiology report generation is its inherent multimodal alignment. Unlike ResNet and vanilla ViT, which must learn image-to-text relationships through supervised report generation loss, CLIP has already been exposed to natural language descriptions during pretraining. As a result, the image embeddings it produces are better aligned with the kinds of features that influence textual generation. This is especially beneficial when using a language model decoder like BioBART, which operates entirely in the textual domain and requires its input to already carry semantic structure.

Moreover, the CLIP encoder enhances lexical diversity and fluency in generated reports. Empirical results in our experiments indicate that pipelines using CLIP as the encoder tend to produce reports that contain more nuanced phrases and domain-specific terminology, even when trained on small datasets like IU-Xray. This improvement stems from the fact that CLIP’s representations encode not just visual structure but also implicit textual associations, learned from its massive image-caption training corpus.

Despite these strengths, CLIP-ViT-B/32 also comes with certain limitations. Firstly, the 32×32 patch size is coarser than that used in ViT-B/16, which may lead to the loss of fine-grained spatial details, particularly relevant in high-resolution medical images where small anomalies can have significant clinical implications. Secondly, CLIP was not pretrained on biomedical data, so its embedding space may not fully reflect the specificity of medical terminology or image semantics. However, this issue is mitigated in our system by using a domain-specific decoder (BioBART), which complements the general-purpose nature of CLIP by grounding the final output in medically accurate language.

As with the other visual encoders in our study, the CLIP encoder is frozen during training. This design choice allows us to preserve the pretrained semantic alignment between vision and language while focusing optimization efforts on the projection module and the language decoder. Freezing also reduces memory usage and training time, making the system more efficient and reproducible. The frozen encoder serves as a fixed feature extractor, producing embeddings that are stable and highly informative for report generation, even under constrained data conditions.

The integration of CLIP into our system reflects a broader trend in machine learning toward foundation models, large pretrained networks that generalize well across tasks with minimal adaptation. By leveraging CLIP’s contrastive pretraining, our framework inherits cross-modal understanding capabilities that are particularly advantageous for bridging the gap between complex radiographic images and diagnostic narratives.

In clinical use cases, this multimodal grounding leads to several practical benefits. First, the generated reports tend to mirror real-world radiological language more closely, including the use of hedging, uncertainty expressions, and comparative findings. Second, the attention weights in the projection and decoder modules, when visualized, reveal that CLIP-based embeddings prompt the decoder to attend to more semantically coherent and anatomically appropriate textual constructions. This contributes to both report coherence and clinical trustworthiness, which are crucial in sensitive applications like diagnostic support systems.

To summarize, CLIP-ViT-B/32 offers a powerful and flexible visual encoding strategy for radiology report generation. Its transformer-based architecture enables global spatial reasoning, while its contrastive multimodal pretraining equips it with semantic awareness that enhances alignment with clinical language. When used in combination with a projection module and a biomedical language model, CLIP-based pipelines deliver high-quality, contextually grounded radiology reports even in low-resource settings. As our experimental results later show, CLIP-ViT-B/32 consistently outperforms ResNet152 and ViT-B/16 in terms of lexical richness, semantic accuracy, and overall fluency, reinforcing its role as a promising visual backbone for future multimodal medical applications.

## MLP Projection Module

One of the most critical yet often overlooked components in multimodal learning systems is the projection module - a bridge that enables effective alignment and communication between modality-specific encoders and decoders. In our framework for chest X-ray report generation, the projection module plays a central role in transforming the visual embedding space produced by the image encoder into the language embedding space expected by the text decoder (BioBART) [16]. This transformation is not merely a dimensional adjustment but a semantic mapping that must preserve clinical relevance, spatial context, and discriminative features while adapting to the sequential structure of language generation.

The projection module in our system is implemented as a multi-layer perceptron (MLP), a feedforward neural network that applies a series of learned affine transformations and nonlinear activations to the input visual features. The choice of MLP as a projection mechanism is motivated by its flexibility, simplicity, and capacity for learning complex transformations without requiring architectural changes to the encoder or decoder. Importantly, by keeping the MLP modular and independent, we enable easy experimentation with different encoders (ResNet, ViT, CLIP) and decoders (BioBART or others), making the system highly extensible and adaptable to new tasks or data typ

At a conceptual level, the MLP projection module receives as input the concatenated image embeddings from the dual-view chest X-rays. Depending on the visual encoder used, this input vector can range in dimensionality from 1024 (CLIP), 1536 (ViT), to 4096 (ResNet). The objective of the MLP is to map this input into a fixed-size latent space that matches the initial input embedding size of the language decoder, which in the case of BioBART is 768 dimensions. This ensures that the decoder can seamlessly condition its generation process on the projected visual features, allowing for fluent and context-aware report generation.

The architecture of the MLP projection module consists of two to three linear layers, each followed by a non-linear activation function (e.g., ReLU or GELU) and dropout for regularization. In our main configuration, we adopt a 2-layer MLP with the following structure:

1. A Linear Layer that reduces the dimensionality from the concatenated feature vector (e.g., 4096) to an intermediate hidden dimension (e.g., 2048).
2. A ReLU activation to introduce non-linearity, enabling the model to learn complex interactions between the features from different views.
3. A second Linear Layer that further projects the hidden representation down to 768 dimensions, matching the input dimensionality of the BioBART decoder.
4. An optional LayerNorm is applied to stabilize training and normalize the output embeddings before feeding them to the decoder.

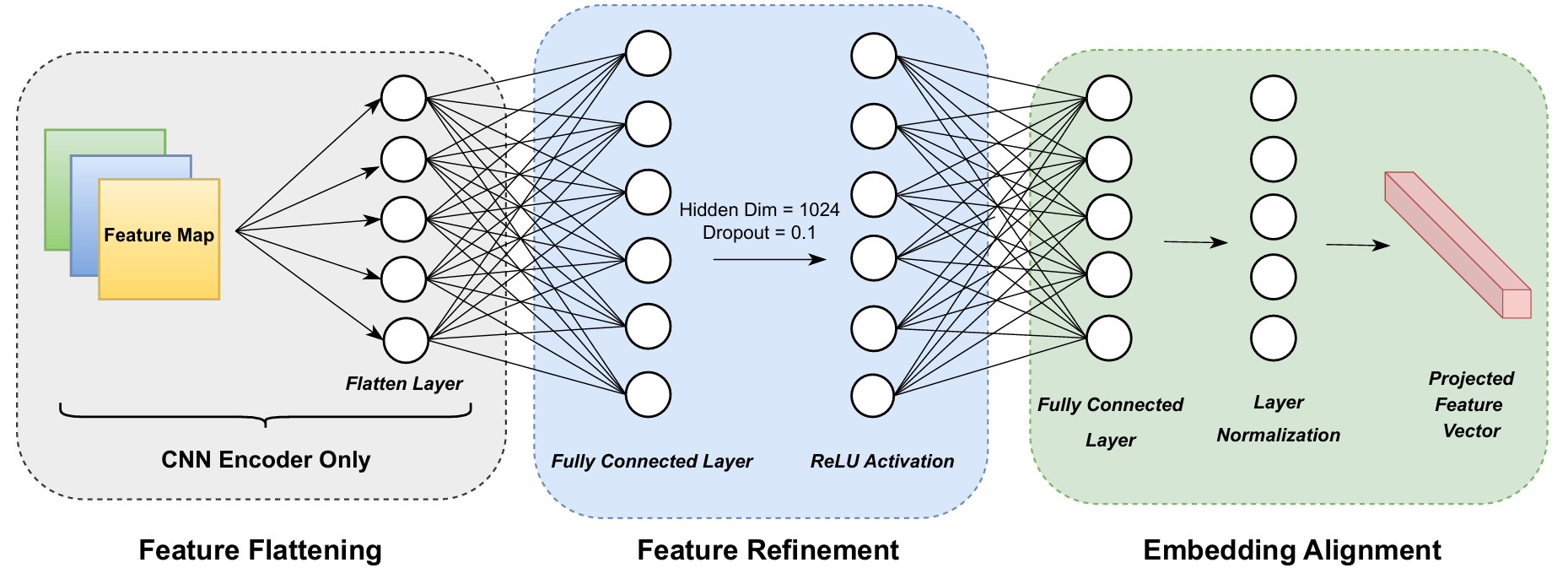


Fig 4. The MLP Architecture for Feature Projecting

This architecture is intentionally shallow. Deeper or wider networks could, in theory, capture more intricate mappings between the image and language domains. However, we found that such designs tend to overfit quickly in the low-data setting of IU-Xray and do not yield consistent performance gains. Empirically, the 2-layer configuration provides a good balance between expressivity and generalization, maintaining stable training dynamics while allowing for sufficient representational power to support high-quality generation.

The design of the MLP module is further motivated by the semantic gap between visual and textual representations [12]. Image embeddings, even when semantically rich, are fundamentally structured as spatially dense, high-dimensional feature vectors, often encoding object boundaries, intensity gradients, and geometric configurations. In contrast, language models like BioBART operate on a sequential stream of token embeddings, where information is organized temporally and hierarchically according to syntactic and semantic structures. The projection module must learn to transform and reframe visual information into a format that is meaningful to the decoder, not just in terms of numerical compatibility, but in terms of semantic alignment and linguistic coherence.

One of the challenges in designing the projection module is ensuring that it preserves clinically relevant information from both images. In the dual-view setup of our system, each image provides a distinct but complementary perspective of the thoracic anatomy. The frontal view is useful for assessing cardiac silhouette, lung fields, and pleural contours, while the lateral view aids in localizing lesions in the anterior-posterior axis and disambiguating overlapping structures. Simply concatenating their embeddings assumes that both contribute equally and independently. However, in practice, some features may be redundant or even contradictory. The MLP is tasked with resolving this redundancy and fusing the information into a single, semantically consistent representation.

To guide the learning of this fusion process, the MLP is trained end-to-end with the decoder. The loss function used during training is a sequence-level cross-entropy loss, computed between the generated report tokens and the ground truth radiology report. Gradients from this loss flow backward through the decoder into the projection module, allowing it to learn how to organize the visual input in a way that best supports accurate text generation. Notably, the visual encoder itself remains frozen, which places a greater learning burden on the MLP. As such, the module must not only bridge the modalities but also extract the most useful subset of the visual information for the task at hand.

A key insight from our experiments is that the MLP does not need to be deep or highly parameterized to be effective. In fact, increasing the depth of the MLP beyond three layers often leads to diminishing returns or even degraded performance. We hypothesize that this is due to a combination of factors: overfitting to the small dataset, increased difficulty in optimization, and the risk of “oversmoothing” the visual features. The simplicity of the 2-layer MLP also contributes to faster convergence, lower memory usage, and improved reproducibility, properties that are particularly desirable in real-world clinical AI systems.

In addition to its core function, the MLP module also serves as a point of control for various experimental interventions. For example, by modifying the activation function, adding normalization layers (LayerNorm or BatchNorm), or changing the output dimension (to match other decoders), we can explore how the nature of the visual-to-language mapping affects report quality. Moreover, by freezing or unfreezing parts of the MLP during training, we can isolate the contribution of learned alignment from the intrinsic power of the encoder or decoder. This makes the projection module not just a bridge but a research probe, capable of revealing insights about cross-modal representation and fusion.

Importantly, the MLP projection is positionally agnostic. Unlike attention-based fusion mechanisms that rely on spatial alignment or patch-wise interactions, our MLP receives global representations already aggregated by the visual encoder (e.g., the [CLS] token from ViT or CLIP, or the global pooling output from ResNet). While this design choice reduces architectural complexity, it does sacrifice fine-grained control over which image regions correspond to specific text tokens. Future work could integrate attention mechanisms that allow the decoder to dynamically attend over image patches or feature maps. However, for the current task, which emphasizes global findings and coherent narrative flow, the projection of a unified global embedding has proven sufficient.

To assess the effectiveness of the MLP module, we conducted an ablation study varying its architectural depth and width. Models with a single linear layer performed noticeably worse, often producing generic or templated reports with limited clinical specificity. In contrast, 2-layer models struck a good balance between accuracy and fluency. Adding a third layer yielded marginal improvements in lexical richness but also increased training time and sensitivity to hyperparameters. These results suggest that model simplicity is not just adequate but advantageous in low-resource medical settings.

Beyond architectural choices, we also evaluated how different encoders interact with the projection module. For instance, CLIP encoders, already aligned with language, required less adaptation and were more tolerant to projection width, while ResNet encoders benefited from slightly larger hidden dimensions. This observation supports the idea that modality alignment is not uniform and that the projection module must be tailored to the characteristics of the encoder.

In conclusion, the MLP projection module is a foundational element of our vision-language framework for automated radiology report generation. By transforming high-dimensional visual features into decoder-compatible embeddings, it enables seamless cross-modal interaction and supports the generation of coherent, clinically meaningful narratives. Its simplicity, flexibility, and efficiency make it well-suited for deployment in real-world healthcare environments, especially when paired with strong pretrained encoders and domain-specific decoders. As vision-language modeling continues to evolve, the projection module remains a key site for innovation, experimentation, and interpretability, with the potential to unlock new frontiers in multimodal clinical AI.

## BioBART-based Report Generation

In the context of automated medical report generation, the language decoder is the final yet arguably the most critical component of the vision-language pipeline. While the visual encoder is responsible for capturing salient features from the radiographic images and the projection module ensures semantic alignment across modalities, the language decoder is tasked with the challenging job of transforming abstract feature representations into coherent, clinically meaningful, and linguistically fluent diagnostic reports. To fulfill this role, we adopt BioBART, a domain-adapted sequence-to-sequence transformer language model that has been pretrained and fine-tuned specifically on biomedical and clinical text corpora.

BioBART is derived from the BART (Bidirectional and Auto-Regressive Transformer) architecture, a denoising autoencoder for pretraining sequence-to-sequence models. BART was originally proposed by Facebook AI and combines the strengths of BERT (bidirectional encoder) and GPT (autoregressive decoder) [17]. In BART, the encoder maps a corrupted version of the input sequence into a latent representation, and the decoder reconstructs the original sequence in an autoregressive manner. This allows BART to serve as a powerful foundation for both text understanding and generation tasks.

BioBART builds upon BART by continuing pretraining and/or fine-tuning on biomedical and clinical corpora, such as PubMed abstracts, PMC full-text articles, and in some versions, clinical notes from MIMIC-III. As a result, BioBART acquires a deep understanding of domain-specific vocabulary, syntactic structures, and semantic conventions used in medical narratives. This makes it particularly well-suited for tasks such as clinical summarization, medical question answering, and radiology report generation, where the language is highly specialized and context-dependent.

1. *Architecture Overview*

BioBART retains the same encoder-decoder architecture as BART, comprising a bidirectional Transformer encoder and an autoregressive Transformer decoder, each consisting of 12 layers (in the base version), multi-head self-attention mechanisms, position embeddings, feed-forward sublayers, and residual connections. However, in our use case, only the decoder portion of BioBART is utilized. This is because the visual modality is already encoded by the image encoder (ResNet, ViT, or CLIP), and the projected embeddings from the MLP module serve as the conditioning input for the decoder.

During inference, the BioBART decoder takes in the projected visual features and generates the report one token at a time, conditioning on the previous tokens and the encoder output (the visual embedding). This autoregressive decoding process continues until an end-of-sequence token is generated or a maximum length threshold is reached. The generation process can be controlled via decoding strategies such as greedy decoding, beam search, top-k sampling, or nucleus sampling, each with trade-offs between determinism, diversity, and computational complexity.

The key strength of using BioBART lies in its ability to generate free-form biomedical text that is both factually grounded and linguistically coherent. Unlike generic language models, which may hallucinate content or misuse terminology in medical contexts, BioBART is explicitly trained on domain-relevant data. This includes specialized phrases, abbreviations, clinical findings, uncertainty expressions, and diagnostic conclusions. Such features are crucial in radiology, where the language used must convey nuanced assessments while maintaining interpretability and clinical validity.

1. *Intergration Into the Pipeline*

In our unified vision-language architecture, the BioBART decoder is integrated seamlessly with the rest of the system. After processing the frontal and lateral chest X-rays through the visual encoder and projecting the concatenated feature vector via the MLP module, the resulting 768-dimensional embedding is fed into BioBART as its initial context. Specifically, the projected vector is reshaped into the position of the first decoder hidden state, effectively serving as a soft prompt that conditions the subsequent token generation process.

Unlike traditional sequence-to-sequence tasks where the encoder processes a source text, our setup treats the image-derived embedding as an abstract semantic representation that triggers the decoder to generate a relevant diagnostic report. This design choice reflects a shift from purely linguistic conditioning to multimodal conditioning, where the decoder must learn to associate semantic features of the input vector such as cardiomegaly, infiltrates, pleural effusion,... with appropriate clinical language in the output.

The decoder uses teacher forcing during training, where the ground-truth tokens are fed into the model one at a time to predict the next token. The loss function is the categorical cross-entropy between the predicted token distribution and the ground-truth token at each position, summed across the sequence. This loss is backpropagated through the decoder and the MLP projection module, but not through the frozen visual encoder. Over time, the model learns to associate certain visual embeddings with the corresponding radiological phrasing, sentence structures, and report flow.

Given the projected visual embedding , we intialize the decoder input embedding as:

|  |  |
| --- | --- |
|  | (4) |

where acts as the initial input to the decoder.

The decoder generates as sequence of token is computed as:

|  |  |
| --- | --- |
|  | (4) |

where is the hidden state at time t from the decoder; and are the output projection matrix and bias, respectively; and Softmax denotes the normalization over the vocabulary.

The overall sequence probability is:

|  |  |
| --- | --- |
|  | (5) |

The model paraments are optimized by minimizing the cross-entropy loss defined as:

|  |  |
| --- | --- |
|  | (6) |

where is the target token at position t.

The model is fine-tuned using paired image-report data, enabling it to learn the mapping from projected visual features to corresponding texual descriptions. Through this approach, BioBART serves as a powerful decoder that translates visual information into precise and medically sound radiology narratives.

1. *Training Details and Optimization*

BioBART is initialized from publicly available pretrained checkpoints, typically trained on PubMed and PMC using span corruption or denoising objectives. In our setting, we fine-tune the decoder on the IU-Xray dataset, which contains 3,955 chest X-ray reports. Each report is preprocessed to extract only the “Findings” section, which serves as the target output. Reports are lowercased, tokenized using the BioBART tokenizer (compatible with SentencePiece), and truncated or padded to a maximum length of 128 tokens.

Fine-tuning is performed using the Adam optimizer with a learning rate of 3e-5 and weight decay of 0.01. We use a batch size of 8 and apply early stopping based on validation loss. Dropout is applied to the decoder layers with a rate of 0.1 to prevent overfitting. The model is trained for up to 20 epochs, although convergence is typically achieved within 10 epochs. Due to the limited size of IU-Xray, we do not perform decoder layer unfreezing or multi-stage fine-tuning; instead, we rely on the pretrained language capabilities of BioBART and the task-specific adaptation provided by the MLP projection.

1. *Advantages in biomedical generationn*

BioBART offers several key advantages for medical report generation:

1. *Domain-aware vocabulary:* BioBART understands medical terminology, enabling it to use appropriate phrasing for findings, anatomical structures, and pathologies.
2. *Fluent language generation:* It produces text that is grammatically correct and stylistically consistent with clinical reports, including hedging, negations, and uncertainty markers.
3. *Generalization from limited data:* Thanks to pretraining, BioBART performs well even when fine-tuned on small datasets like IU-Xray, avoiding the need for extensive annotation.
4. *Interpretability:* Attention maps in the decoder can be visualized to understand how the model relates specific visual features to tokens in the report.
5. *Compatibility:* Being a transformer model, BioBART integrates smoothly with transformer-based encoders and supports further extensions such as prefix tuning, adapter layers, or prompt-based conditioning.
6. *Limitation, challenges and future directions*

Despite its advantages, BioBART is not without limitations. One of the primary challenges is the semantic grounding of generated text. Since the decoder is pretrained on textual data and receives only a single projected embedding from the visual encoder, it may sometimes produce plausible-sounding but factually incorrect statements, especially when visual cues are ambiguous. This issue is compounded by the lack of explicit attention to image regions, as the MLP projection does not retain spatial resolution.

Another challenge lies in report structure. While BioBART can generate fluent sentences, it may struggle to follow the structured logic of radiological reporting, such as maintaining the anatomical order or relating multiple findings within the same region. Incorporating explicit structural priors or hierarchical generation techniques could address this issue.

Furthermore, BioBART’s decoder length limitation (128 tokens) may restrict its ability to generate longer or more detailed reports, especially for complex cases. Extending the maximum sequence length or adopting hierarchical decoding could mitigate this constraint.

BioBART opens up exciting directions for future research. One avenue is to augment the decoder with cross-attention to visual patches, enabling token-level alignment between words and image regions. This would increase interpretability and grounding. Another is to incorporate structured output formats, such as section-wise generation or template-based prompting, to better mimic real-world reporting practices.

Additionally, integrating knowledge graphs or medical ontologies into the decoder could further enhance factual consistency and reduce hallucinations. Finally, training BioBART on multi-institutional datasets or combining it with multilingual corpora could improve its robustness and generalizability in diverse clinical settings.

# Experiments

1. Experimental Setup

Our experiments are conducted on the IU-Xray dataset [18], a benchmark corpus for medical report generation that consists of 3,955 radiology reports and 7,470 chest X-ray images. Each report is associated with one or two images, corresponding to frontal and lateral views of the patient. For the purposes of our study, we select only those samples where both views are available, resulting in a filtered dataset of 2,750 training, 563 validation, and 394 test samples.

Each report is preprocessed to extract the “Findings” section, which serves as the generation target. The average report length is approximately 90 words, or 45–60 tokens. We tokenize the text using the BioBART tokenizer, and truncate sequences longer than 128 tokens. Image preprocessing includes resizing to 224×224 pixels and normalization. Data augmentation such as horizontal flipping and rotation is applied during training but not during evaluation.

The model is implemented in PyTorch, using the HuggingFace Transformers library for loading and fine-tuning BioBART. All visual encoders (ResNet152, ViT-B/16, CLIP-ViT-B/32) are loaded from pretrained weights and frozen during training. Only the MLP projection module and the BioBART decoder are trained.

1. Training Configruation

Training is performed using the Adam optimizer with a learning rate of 3e-5 and a weight decay of 0.01. A batch size of 8 is used, with gradient accumulation over two steps to simulate a batch size of 16. Dropout with a probability of 0.1 is applied to the projection module and the decoder layers to prevent overfitting. The model is trained for 20 epochs, with early stopping triggered if the validation loss does not improve over 5 consecutive epochs.

1. Evaluation

Model performance. Table I summarizes the performance of various pipelines evaluated using ROUGE [22], BLEU [23], and METEOR [24] metrics. Among the models, our proposed CLIP–MLP–BioBART configuration achieves the highest scores in BLEU-1 (0.480), ROUGE (0.3986), and METEOR (0.4175), indicating stronger overall performance in terms of fluency, semantic coverage, and alignment. These results suggest that the combination of CLIP’s global visual understanding and BioBART’s domain-adapted text generation contributes to improved lexical diversity and conceptual accuracy in the generated reports.

In contrast, the ResNet152–recurrent–attention pipeline outperforms other configurations in terms of BLEU-2 (0.358), BLEU-3 (0.270), and BLEU-4 (0.195). This suggests that recurrent CNN-based models, while more limited in capturing global context, may have an advantage in preserving short and mid-length n-gram patterns, possibly due to their memorization capacity and inductive bias toward local sequential structures.

These findings highlight the strength of transformer-based encoders, particularly CLIP, in aligning visual inputs with textual outputs at a conceptual level. However, the relatively lower BLEU-2 to BLEU-4 scores in the CLIP-based model may reflect limitations in reproducing exact n-gram patterns, especially under a low-resource setting where the BioBART decoder may not be fully adapted to the distribution of domain-specific phrase structures.

Tab 1. Performance Comparision of each pipeline on the test set

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Pipeline method | Rouge-L | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | Meteor |
| Vanilla CNN-RNN [4] | 0.226 | 0.273 | 0.144 | 0.116 | 0.082 | 0.125 |
| Hierachical generation [5] | 0.325 | 0.437 | 0.323 | 0.221 | 0.172 | 0.244 |
| Resnet152-recurrent-conv [6] | 0.309 | 0.416 | 0.298 | 0.217 | 0.163 | 0.227 |
| Resnet152-recurrent-BiLSTM [6] | 0.322 | 0.423 | 0.307 | 0.223 | 0.165 | 0.236 |
| Resnet152-recurrent-attention [6] | 0.366 | 0.464 | 0.358 | 0.270 | 0.195 | 0.274 |
| (Ours) Resnet152 - MLP - BioBART | 0.375 | 0.426 | 0.272 | 0.179 | 0.124 | 0.379 |
| (Ours) ViT-B/16-MLPBioBART | 0.410 | 0.465 | 0.297 | 0.188 | 0.123 | 0.395 |
| (Ours) CLIP - MLP - BioBART | 0.399 | 0.480 | 0.311 | 0.206 | 0.136 | 0.418 |

Despite achieving notable improvements across all metrics when compared to prior baselines, the proposed pipeline is not yet ready for clinical deployment. As illustrated in Figure 5 (a) and (b), the generated reports mostly align with the reference texts, capturing general findings and report structure accurately. However, some clinically relevant details are either omitted or incompletely represented in both the generated and reference reports, even when no direct contradictions are observed.

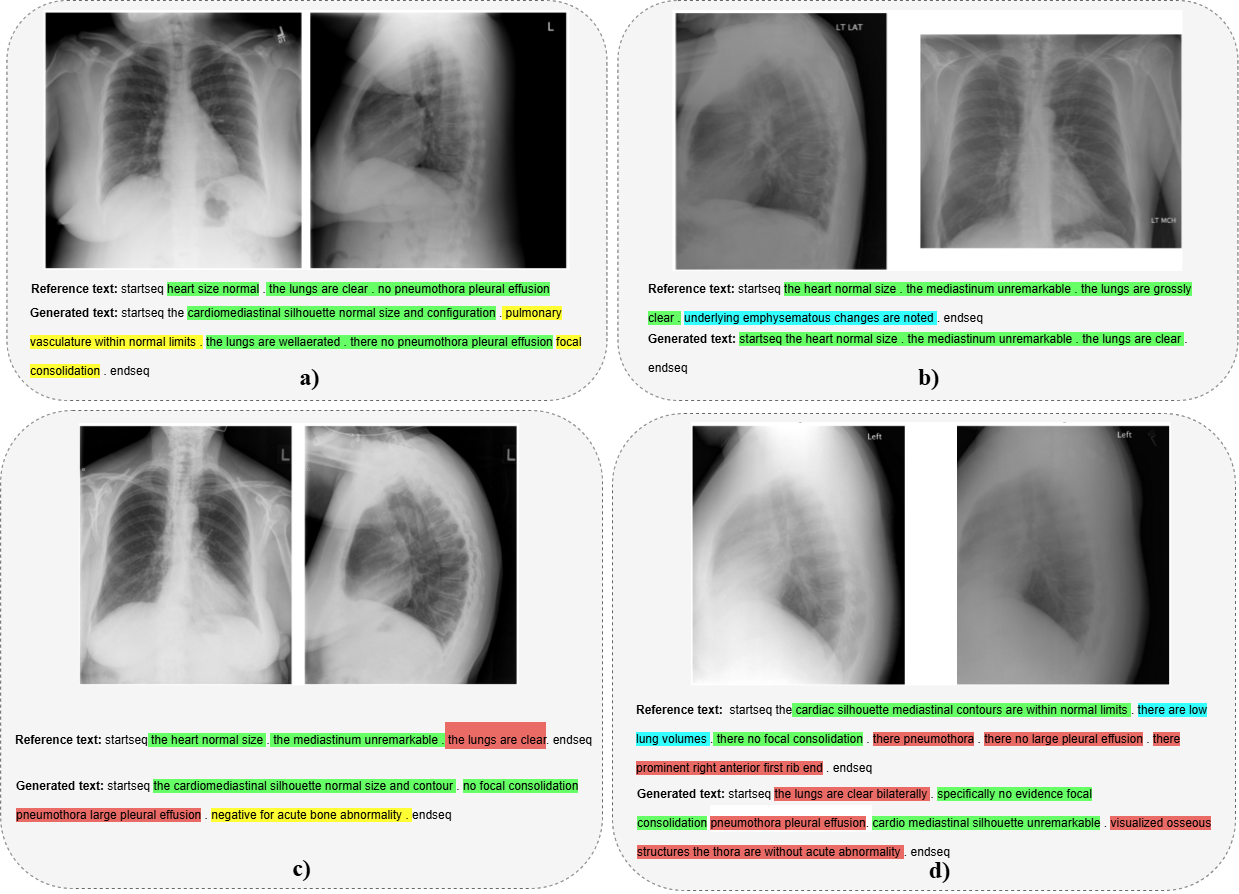


Fig 5. Comparison between generated reports and its corresponding reference report from the test set. Text segments with identical or semantically equiv- alent meanings in both the generated and reference reports are highlighted in green. Contradictory content between the two reports is marked in red. Cyan highlights information present only in the reference report, while yellow highlights content exclusive to the generated report.

More concerning are the cases shown in Figure 5 (c) and (d), where the generated content only partially aligns with the reference, primarily by reproducing common boilerplate phrases seen frequently in the dataset. This form of lexical agreement, while contributing to high BLEU or ROUGE scores, lacks discriminative value in a clinical context. Specifically, Figure 5(d) presents a semantic contradiction: while the reference report describes clear lung abnormalities, the generated report erroneously states that no significant findings are present. Such mismatches highlight the critical importance of clinical factuality and the challenges associated with training under limited data conditions [25].

These observations reinforce the need for more robust multimodal alignment mechanisms, improved domain adaptation, and error-sensitive training objectives to ensure that high-scoring outputs are not only linguistically coherent but also clinically trustworthy.

1. Ablation Study

To examine the influence of architectural complexity within the projection component of our framework, we conducted an ablation study focusing on the MLP projection module, specifically its internal Feature Refinement Block design. Our objective was to investigate how varying the depth of this component would affect the overall performance and training stability of the system. To this end, we evaluated three configurations of the CLIP–MLP–BioBART pipeline, characterized by the number of sequential refinement layers used within the MLP: a 1-block, 2-block, and 3-block architecture.

In the baseline configuration, the projection module consists of a single linear transformation followed by a ReLU activation, forming one Feature Refinement Block. The modified variants extend this by stacking two and three such blocks sequentially. For each architectural setting, we performed 10 independent training runs with identical hyperparameters to measure both average performance and standard deviation, providing insight into model robustness and convergence consistency under different levels of MLP depth.

Tab 2. Model average performance of MLP settings across 10 runs

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| MLP Settings | Rouge-L | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | Meteor |
| Feature Refinement block x 1 | 0.399 | 0.480 | 0.311 | 0.206 | 0.136 | 0.418 |
| Feature Refinement block x 2 | 0.363 | 0.351 | 0.219 | 0.139 | 0.091 | 0.333 |
| Feature Refinement block x 3 | 0.367 | 0.341 | 0.215 | 0.137 | 0.0898 | 0.332 |

Tab 3. Model average performance of MLP settings across 10 runs

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| MLP Settings | Rouge-L | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | Meteor |
| Feature Refinement block x 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Feature Refinement block x 2 | 0.016 | 0.098 | 0.063 | 0.040 | 0.024 | 0.0498 |
| Feature Refinement block x 3 | 0.029 | 0.102 | 0.070 | 0.048 | 0.032 | 0.062 |

Tables II and III summarize the results of this evaluation. Table II presents the mean scores across the BLEU, ROUGE, and METEOR metrics, while Table III details the standard deviation for each score across the 10 runs. The results clearly indicate that the 1-block configuration consistently achieves the best average performance across all metrics. Furthermore, it demonstrates zero standard deviation, implying that model behavior is stable and reproducible across runs-a highly desirable property in medical applications.

In contrast, the 2-block configuration, while still competitive in terms of ROUGE and METEOR, exhibits increased variability. Although it marginally improves METEOR, it performs slightly worse on the BLEU scores, suggesting that while deeper MLPs may help improve semantic alignment, they may not necessarily preserve lexical precision or n-gram consistency. The 3-block configuration, despite its increased representational capacity, fails to deliver consistent performance improvements and demonstrates the highest standard deviation, indicative of training instability. This configuration is also more prone to converge to suboptimal local minima, particularly under low-resource training conditions such as the IU-Xray dataset.

These findings highlight a critical trade-off between model complexity and training stability. While deeper architectures might theoretically allow for richer mappings between visual and linguistic embeddings, they are also more sensitive to overfitting, learning rate fluctuations, and gradient vanishing. In practice, our results suggest that a compact 1-block MLP architecture strikes the optimal balance, offering not only strong performance but also robustness to initialization and stochastic training effects.

In summary, the ablation study reveals that increasing the complexity of the projection module does not necessarily translate to improved model performance. On the contrary, under limited training data conditions, simpler architectures prove to be both more effective and more reliable, supporting the design decision to adopt a 2-layer MLP with a single refinement block in our final model.

# Conclusion

In this study, we introduced a novel multimodal framework for automated medical imaging report generation, combining advanced visual feature extraction with domain-adapted natural language generation. Our architecture leverages recent advances in vision-language modeling by integrating state-of-the-art visual encoders, namely Vision Transformers (ViT-B/16) and CLIP-ViT-B/32 with BioBART, a pretrained biomedical sequence-to-sequence language model, to generate diagnostic narratives from chest X-ray images. To enable effective communication between the visual and textual modalities, we proposed a lightweight yet effective MLP-based projection module that aligns visual embeddings into the semantic space expected by the language decoder.

The core strength of our system lies in its ability to generate clinically meaningful, fluent, and structurally coherent reports using only a limited amount of annotated training data. Through extensive experiments on the IU-Xray benchmark dataset, we demonstrated that transformer-based vision encoders consistently outperform conventional CNN backbones ( ResNet152), particularly in capturing global spatial patterns and producing richer semantic representations. In parallel, the use of domain-specific language modeling via BioBART significantly improved the clinical accuracy and fluency of the generated text compared to general-purpose text decoders.

Our ablation studies further highlighted the importance of architectural simplicity in the projection module. Increasing the depth of the MLP beyond a single or two-block configuration resulted in diminished performance stability and higher variance, especially under data-scarce conditions. These findings reinforce the notion that compact, well-regularized designs often offer a more reliable solution in real-world clinical NLP settings where large-scale labeled datasets are not always available.

While the model performs well on average and offers clear advantages in both quantitative and qualitative evaluations, certain limitations remain. Notably, the model may struggle with rare pathologies, generate generic phrasing, or underperform in token-level grounding due to the lack of explicit attention mechanisms between image regions and text tokens. Future work could address these limitations through techniques such as cross-attention-based fusion, entity-aware decoding, or incorporation of structured clinical knowledge graphs.

In conclusion, this research contributes to a scalable, extensible, and effective vision-language generation framework tailored for the biomedical domain. We hope our work provides a solid foundation for future efforts that seek to harness large pretrained models for clinical applications, especially in scenarios constrained by limited data availability. More broadly, we envision this line of research contributing to the advancement of AI-assisted clinical documentation, ultimately improving diagnostic efficiency, patient safety, and healthcare accessibility across diverse global contexts.

# References

1. B. Jing, P. Xie, and E. Xing, “On the automatic generation of medical imaging reports,” *arXiv preprint arXiv:1711.08195*, 2017.
2. A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is all you need,” Advances in Neural Information Processing Systems, vol. 30, 2017.
3. Y. Nie, L. Li, Z. Gan, et al., “Mlp architectures for vision- and-language modeling: An empirical study,” arXiv preprint arXiv:2112.04453, 2021.
4. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.
5. J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, “A hierar- chical approach for generating descriptive image paragraphs,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 317–325.
6. Y. Xue, T. Xu, L. Rodney Long, et al., “Multimodal recurrent model with attention for automated radiology report genera- tion,” in Medical Image Computing and Computer Assisted Intervention–MICCAI 2018, 2018, pp. 457–466.
7. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint arXiv:2010.11929, 2020.
8. S. Regmi, A. Subedi, U. Bagci, and D. Jha, “Vision trans- former for efficient chest x-ray and gastrointestinal image classification,” arXiv preprint arXiv:2304.11529, 2023.
9. A. Radford, J. W. Kim, C. Hallacy, et al., “Learning trans- ferable visual models from natural language supervision,” in International Conference on Machine Learning, 2021, pp. 8748–8763.
10. H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu, “Biobart: Pretraining and evaluation of a biomedical generative language model,” arXiv preprint arXiv:2204.03905, 2022.
11. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” OpenAI blog, vol. 1, no. 8, p. 9, 2019.
12. R. Mokady, A. Hertz, and A. H. Bermano, “Clipcap: Clip prefix for image captioning,” arXiv preprint arXiv:2111.09734, 2021.
13. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in CVPR, 2016, pp. 770–778.
14. V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.
15. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” The Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.
16. J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” arXiv preprint arXiv:1607.06450, 2016.
17. M. Lewis, Y. Liu, N. Goyal, et al., “Bart: Denoising sequence- to-sequence pre-training for natural language generation, trans- lation, and comprehension,” arXiv preprint arXiv:1910.13461, 2019.
18. D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, et al., “Preparing a collection of radiology examinations for distribution and retrieval,” J. Am. Med. Inform. Assoc., vol. 23, no. 2, pp. 304–310, 2015.
19. I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in Advances in Neural Infor- mation Processing Systems, vol. 27, 2014.
20. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.
21. L. Prechelt, “Early stopping-but when?” In Neural Networks: Tricks of the Trade, Springer, 2002, pp. 55–69.
22. K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Jul. 2002, pp. 311–318.
23. C. Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Text Summarization Branches Out, Jul. 2004, pp. 74–81.
24. S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in Proceedings of the ACL Workshop on Intrinsic and Ex- trinsic Evaluation Measures for Machine Translation and/or Summarization, Jun. 2005, pp. 65–72.
25. W. Boag, E. Sergeeva, S. Kulshreshtha, P. Szolovits, A. Rumshisky, and T. Naumann, “Cliner 2.0: Accessible and accurate clinical concept extraction,” arXiv preprint arXiv:1803.02245, 2018.