Med3DVLM: An Efficient Vision-Language Model for 3D Medical Image Analysis

Yu Xin[1†], Gorkem Can Ates[1†], Kuang Gong[1], Wei Shao[1*]

[1]University of Florida

[†]Contribute equally [*]Corresponding to: weishao@ufl.edu

Abstract

Vision-language models (VLMs) have shown promise in 2D medical image analysis, but extending them to 3D remains challenging due to the high computational demands of volumetric data and the difficulty of aligning 3D spatial features with clinical text. We present Med3DVLM, a 3D VLM designed to address these challenges through three key innovations: (1) DCFormer, an efficient encoder that uses decomposed 3D convolutions to capture fine-grained spatial features at scale; (2) SigLIP, a contrastive learning strategy with pairwise sigmoid loss that improves image-text alignment without relying on large negative batches; and (3) a dual-stream MLP-Mixer projector that fuses low- and high-level image features with text embeddings for richer multi-modal representations. We evaluated our model on the M3D dataset, which includes radiology reports and VQA data for 120,084 3D medical images. The results show that Med3DVLM achieves superior performance on multiple benchmarks. For image-text retrieval, it reaches 61.00% R@1 on 2,000 samples, significantly outperforming the current state-of-the-art M3D-LaMed model (19.10%). For report generation, it achieves a METEOR score of 36.42% (vs. 14.38%). In open-ended visual question answering (VQA), it scores 36.76% METEOR (vs. 33.58%), and in closed-ended VQA, it achieves 79.95% accuracy (vs. 75.78%). These results demonstrate Med3DVLM's ability to bridge the gap between 3D imaging and language, enabling scalable, multi-task reasoning across clinical applications. Our code is publicly available at https://github.com/mirthAI/Med3DVLM.

1. Introduction

Medical image analysis plays a crucial role in diagnosing and treating diseases such as cancer, cardiovascular conditions, and neurological disorders. However, existing models are often task-specific, lack adaptability to new tasks, and do not support real-time user interactions. Vision-language models (VLMs), such as Contrastive Language-Image Pre-Training (CLIP) [28] and Large Language and Vision Assistant (LLaVA) [20], offer greater versatility in dynamic clinical settings by aligning medical images with textual reports. CLIP leverages contrastive learning on image-text pairs to enable zero-shot image classification and image-text retrieval. LLaVA extends CLIP by integrating its visual encoder with a large language model (LLM), facilitating interactive tasks such as report generation and visual question answering (VQA). Despite their success in analyzing 2D medical images, such as chest X-rays [8] and pathology slides [22], their application to 3D imaging remains limited.

3D imaging modalities, such as computed tomography (CT) and magnetic resonance imaging (MRI), provide volumetric data that capture spatial details unavailable in 2D images. However, extending VLMs to 3D presents significant challenges. First, a 3D scan consists of hundreds of slices, making slice-by-slice analysis prone to losing global context. Second, directly building 3D VLMs significantly increases computational complexity due to the higher dimensionality. Third, the scarcity of publicly available 3D image-report pairs further limits model development.

Several VLMs have attempted to bridge this gap, but each has limitations in the 3D setting. For example, PMC-CLIP [18], trained on large-scale biomedical literature images, is restricted to 2D inputs, leading to poor performance on 3D image understanding. RadFM [39] unifies 2D and 3D data but is primarily optimized for text generation tasks, such as VQA, and struggles with broader image-text understanding. More recently, M3D-LaMed [2] was introduced as a generalist multi-modal model for 3D medical image analysis. By combining a CLIP-pretrained 3D visual encoder with a 3D spatial pooling perceiver, M3D-LaMed enables direct reasoning on 3D scans and achieves state-of-the-art performance across multiple benchmarks, including image-text retrieval, report generation, open-ended and

closed-ended VQA, as well as 3D segmentation and localization tasks.

Despite these advancements, M3D-LaMed has notable limitations. Its 3D backbone, similar to many other 3D VLMs, incurs high computational costs when processing high-resolution 3D volumes, as standard 3D vision transformers scale poorly with image size. Additionally, M3D-LaMed's vision-language alignment relies on CLIP's contrastive loss, which compares each image-text pair against a large set of negative pairs in the batch. This approach works well for large datasets but can be less effective in smaller medical datasets, where meaningful negative samples are limited, and similar images may still share semantic information. Furthermore, its multi-modal fusion mechanism, which projects visual features into the LLM via a multilayer perceptron (MLP), may not sufficiently capture complex cross-modal interactions.

To overcome the limitations of existing VLMs in 3D imaging, we introduce Med3DVLM, a novel VLM that incorporates three key innovations to improve 3D feature learning, cross-modal alignment, and multi-modal projection:

1. **Efficient 3D Feature Encoding.** We integrate DCFormer [1], a 3D image encoder that efficiently captures volumetric features by decomposing 3D convolutions into three parallel 1D convolutions along the depth, height, and width axes. This approach reduces computational complexity and enables richer, more scalable feature representations for 3D image volumes.

2. **Improved Vision-Language Alignment.** We adopt SigLIP [45], a sigmoid-based language-image pretraining scheme, to improve image-text alignment. Unlike CLIP's softmax-based contrastive loss, which relies on distinguishing positive pairs from a large batch of negative samples, SigLIP uses a pairwise sigmoid loss that directly optimizes each image-text pair independently. This eliminates the need for global similarity normalization across batches, making training more stable and less sensitive to batch size.

3. **Multi-Scale Multi-Modal Projector.** We introduce a novel projector based on MLP-Mixer [30] to fuse image and text embeddings effectively. Utilizing a low-high-hybrid design, it blends detailed low-level and abstract high-level features from the image encoder with LLM text embeddings using stacked MLP layers. Inspired by MLP-Mixer's ability to mix spatial and feature data, this dual-stream approach captures richer cross-modal interactions than simple linear projection, improving the LLM's decoding accuracy.

2. Related Work

2.1. Medical Vision-Language Models

Medical VLMs utilize multi-modal learning to enhance tasks such as disease diagnosis, report generation, and medical VQA through improved integration and understanding of medical images and text. One of the earliest efforts, ConVIRT [47], applied contrastive learning to learn visual representations from paried image-text data, outperforming ImageNet pretraining in medical image classification and zero-shot retrieval tasks. Similarly, BioViL [4] leveraged large-scale biomedical datasets to refine multi-modal representations, significantly outperforming previous supervised methods. Further extending these capabilities, MedCLIP [36] replaced the InfoNCE loss with semantic matching loss based on medical knowledge, demonstrating its ability to learn generalized representations with limited data.

Recent efforts have aimed to develop VLMs for 3D medical imaging. RadFM [39] introduced a generalist VLM trained on massive datasets that consisted of both 2D and 3D medical images and associated radiology reports. It integrated contrastive learning and generative modeling across diverse imaging modalities (X-ray, CT, MRI) and textual reports, enabling a unified radiology representation. Extending this, M3D [2] introduced a large-scale dataset, M3D-Data, along with M3D-LaMed, a multi-modal vision-language model for 3D medical image analysis tasks such as image-text retrieval, report generation, VQA, and promotable segmentation. Its successor, E3D-GPT [15], employed a 3D multi-modal masked autoencoder framework to further enhance image-text retrieval, leading to improved performance in report generation and VQA. In parallel, researchers have also explored text-guided 3D medical image segmentation [48, 14, 40], where language prompts are used to localize anatomical structures or pathologies.

2.2. Radiology Report Generation

Automated report generation aims to produce descriptive, accurate, and clinically relevant reports from medical images. It improves diagnostic accuracy while reducing radiologists' workload and supporting care in resource-limited settings. Early approaches primarily adopted an encoder-decoder architecture, where convolutional neural networks (CNNs) extracted image features, and long short-term memory (LSTM) networks generated text descriptions [41, 12, 44, 37]. However, these methods struggled with long-range dependencies, often produced repetitive text, and had limited capacity to capture complex medical semantics.

The transformer architecture [33] addressed key limitations of CNNs and LSTMs by using self-attention to capture global dependencies. Pretrained LLMs [31, 27]

based on transformers are now adapted into medical VLMs for report generation. R2GenGPT [35] introduced an efficient visual alignment module to better integrate image features with LLM word embeddings, improving text coherence and clinical relevance. Med-Flamingo [23], based on OpenFlamingo-9B, was pretrained on interleaved medical image-report pairs, achieving superior performance in generating clinically useful responses. More recent methods, such as LLaVA-Med, were based on the LLaVA [20] framework, which combined a visual encoder with an LLM to generate detailed and coherent reports. CT2Rep [11] proposed an advanced 3D vision encoder to generate radiology reports for 3D medical imaging, specifically targeting chest CT volumes. Similarly, CT-CHAT [10] adapted the LLaVA framework for chest CT report generation, demonstrating the effectiveness of large-scale pretrained LLMs in 3D medical imaging.

2.3. Medical Visual Question Answering

A medical VQA system can answer natural language questions about medical images. This task can be closed-ended (with answers like yes/no or a choice from a fixed list) or open-ended (free-form text answers). Early work on medical VQA largely treated it as a classification problem, where models selected the correct answer from a predefined set of possible responses [24, 49, 9, 19]. These models typically encoded images (e.g., using CNNs) and text (e.g., using transformers), and then mapped the combined features to a predefined answer space. Although this approach worked well for simple questions, such as identifying an organ, it struggled with questions requiring detailed explanation or answers not included in the predefined list.

Nowadays, approaches to medical VQA increasingly adopt generative architectures, treating answer generation as a sequence prediction task, which enables open-ended responses. An early step in this direction was CGMVQA [29], which added a generative decoder branch alongside the traditional classifier. Recent VQA systems have fully transitioned to generative models [7]. These models (e.g., Med-PaLM M [32] and LLaVA-Med [16]) extend an LLM with visual inputs via visual instruction tuning [20], allowing the LLM to interpret medical images and answer questions in a conversational manner. Recently, a few works have explored VQA on 3D medical images [2, 10], where models can directly perceive entire 3D image volumes and generate context-aware responses, offering more accurate and holistic medical insights.

3. Methods

3.1. Dataset

This study used the publicly available M3D dataset [2], collected with informed consent and ethical approval by the original investigators. All data were de-identified, and no additional consent or approval was required for this secondary analysis. We used two subsets of the M3D dataset: M3D-Cap and M3D-VQA. M3D-Cap includes 120K image-text pairs, while M3D-VQA contains 662K instruction-response pairs. The dataset can support tasks such as image-text retrieval, radiology report generation, and VQA. All image volumes were resampled to a fixed size of 128x256×256. The dataset was divided into training (115k samples), validation (3k samples), and test (2k samples) sets.

3.2. Med3DVLM

We introduce Med3DVLM, a 3D medical VLM consisting of three core components: a vision encoder, a multi-modal projector, and a large language model. The vision encoder extracts detailed visual features from 3D medical image volumes. These features are integrated with text embeddings through the multi-modal projector, facilitating cross-modal interactions. The LLM then generates coherent, contextually accurate outputs based on these fused features. The overall model architecture is illustrated in Figure 1. The training process is divided into three stages: (1) contrastive pretraining, (2) multi-modal projector pretraining, and (3) VLM fine-tuning.

3.2.1 Contrastive Pretraining

Following the CLIP framework [28], the vision and text encoders are trained jointly using a contrastive loss to align 3D images with their corresponding radiology reports. The goal is to maximize the similarity between embeddings of positive (matching) image-text pairs while minimizing it for negative (non-matching) pairs.

For the vision encoder, we utilize DCFormer [1], which efficiently extracts volumetric features using decomposed 3D convolutions. This approach mitigates computational challenges in high-resolution 3D image analysis, where traditional 3D CNNs and vision transformers struggle due to the cubic scaling of 3D CNNs with kernel size and the quadratic scaling of vision transformers with image size. DCFormer decomposes 3D convolutions into three 1D parallel depthwise convolutions:

$$X_{\mathrm{h}}^{*} = \mathrm{DWConv}_{C\to C}^{k_h\times 1\times 1}(X), \tag{1}$$

$$X_{\mathrm{w}}^{*} = \mathrm{DWConv}_{C\to C}^{1\times k_w\times 1}(X), \tag{2}$$

$$X_{\mathrm{d}}^{*} = \mathrm{DWConv}_{C\to C}^{1\times 1\times k_d}(X), \tag{3}$$

where $X \in \mathbb{R}^{B\times C\times H\times W\times D}$ denote the input feature map, B is the batch size, C is the number of channels, and H, W, and D represent the spatial dimensions (height, width, and depth), respectively. The kernel sizes in these dimensions, denoted as (k_h, k_w, k_d), are set by default to

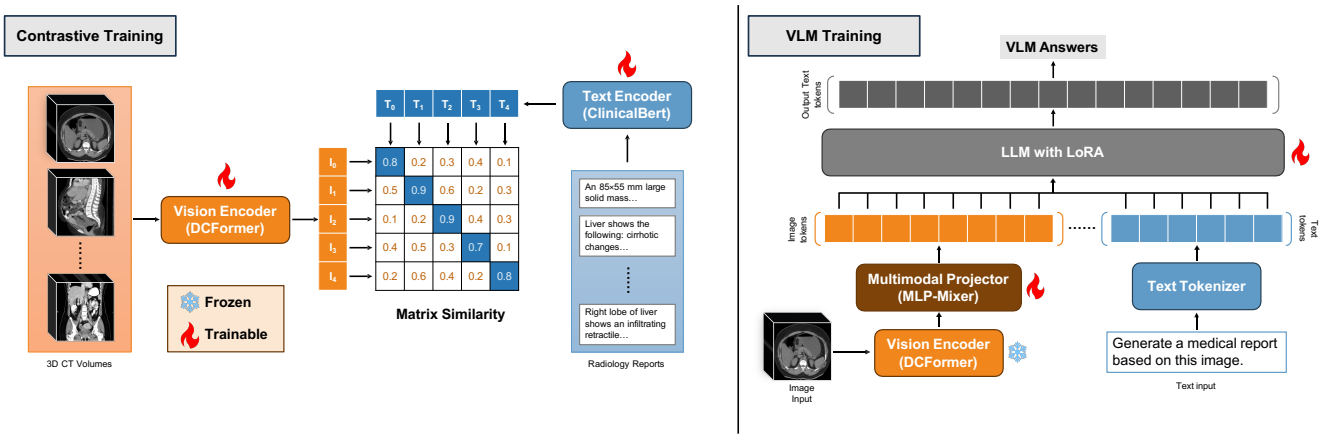


Figure 1: Overview of our Med3DVLM model.

$k_h = k_w = k_d = k \in \{13, 11, 9, 7\}$ to leverage large kernels for improved feature extraction. The extracted features are then normalized and summed to produce the final output:

$$X' = X + \text{Norm}_h(X^*_{\text{h}}) + \text{Norm}_w(X^*_{\text{w}}) + \text{Norm}_d(X^*_{\text{d}}). \tag{4}$$

Such a decomposition strategy enables the processing of 3D volumes at a larger size of 128x256×256 in this study, preserving fine-grained spatial details for improved image-text alignment. DCFormer, short for DeComposed Former, derives its name from this decomposition strategy. The "Former" suffix is inspired by the MetaFormer [43] framework, which generalizes Transformer-like architectures by replacing attention mechanisms with alternative token-mixing operations.

In this work, we utilize DCFormer-small variant, which consists of a stem stage and four hiararchical stages with [2, 3, 6, 2] layers and [96, 192, 384, 768] channels [1]. This adaptation of DCFormer within the LLaVA [20] framework enables efficient processing of high-resolution 3D volumes while preserving spatial detail. By reducing computational complexity, it supports scalable 3D medical vision-language modeling under typical GPU constraints and addresses key limitations of 3D vision transformers and CNNs.

For the text encoder, we use ClinicalBERT [34], an adaptation of BERT [6] pretrained on clinical notes. Designed for clinical text, ClinicalBERT is well-suited for processing radiology reports, generating embeddings that effectively capture their semantic content.

Traditional softmax-based contrastive learning (e.g., CLIP) distinguishes positive pairs from a large batch of negative samples. While effective for large datasets, this approach struggles with smaller medical datasets, where meaningful negative samples are limited, and similar images may share semantic information. To address this, we adopt SigLIP [45], a sigmoid-based language-image pre-training scheme that directly optimizes each image-text pair independently. This eliminates the need for global similarity normalization across batches, resulting in more stable training and reduced sensitivity to batch size. Sigmoid loss is defined as:

$$\mathcal{L}_{\text{SigLIP}} = -\frac{1}{|B|}\sum_{i=1}^{|B|}\sum_{j=1}^{|\mathcal{B}|}\log \underbrace{\frac{1}{1 + e^{z_{ij}(-t\mathbf{x}_i \cdot \mathbf{y}_j + b)}}}_{\mathcal{L}_{ij}}, \tag{5}$$

where z_{ij} is the label for the i-th image and j-th text pair, $z_{ij} = 1$ for positive pairs and $z_{ij} = 0$ otherwise. The sigmoid loss $\mathcal{L}_{ij}$ is computed for each image-text pair, with t as the temperature parameter, $\mathbf{x}_i$ and $\mathbf{y}_j$ as the image and text embeddings, respectively, and b as a bias term. The overall loss is averaged over all pairs in the batch.

For comparison, the CLIP loss is given by:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\underbrace{\log \frac{e^{t\mathbf{x}_i \cdot \mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|} e^{t\mathbf{x}_i \cdot \mathbf{y}_j}}}_{\text{image}\to\text{text softmax}} + \underbrace{\log \frac{e^{t\mathbf{x}_i \cdot \mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|} e^{t\mathbf{x}_j \cdot \mathbf{y}_i}}}_{\text{text}\to\text{image softmax}}\right) \tag{6}$$

Unlike traditional medical VLMs that rely on CLIP pretraining and require large batch sizes for effective contrastive learning, SigLIP adopts a pairwise sigmoid loss that operates independently on each image-text pair. Rather than comparing one positive against many negatives using softmax, SigLIP formulates contrastive learning as a binary classification problem. Matched pairs are pushed together, and unmatched pairs are pushed apart. This design removes the need for batch-wide normalization, improves alignment stability, and reduces sensitivity to batch size, which makes

it more suitable for small-batch, semantically rich medical imaging settings.

3.2.2 Multi-modal Projector Pretraining

We propose an effective multi-modal projector for the interaction between visual and textual embeddings. In this pretraining stage, all model weights are frozen except for those in the projector. To effectively capture spatial and feature dependencies, we adopt the MLP-Mixer architecture [30], which alternates between a token-mixing MLP (spatial mixer) and a channel-mixing MLP.

Given an input image feature map $X \in \mathbb{R}^{n \times d}$, where n is the number of tokens and d is the embedding dimension, the MLP-Mixer operations are defined as:

$$U = W_2\, \sigma(W_1\, (\mathrm{Norm}(X)^\top)), \tag{7}$$

$$Y = W_4\, \sigma(W_3\, \mathrm{Norm}(U^\top)), \tag{8}$$

where W_1, W_2, W_3, W_4 are learnable weight matrices, and σ denotes a nonlinear activation function (e.g., GELU). The token-mixing MLP captures interactions across spatial tokens, while the channel-mixing MLP models dependencies across feature channels.

To further enhance multi-modal fusion, we propose a low-high hybrid Mixer-MLP architecture (Figure 2), inspired by Janus Pro [5]. Unlike M3D-LaMed [2], which produces large spatial outputs $(B, 2048, 768)$ that require lossy downsampling, DCFormer outputs semantically rich features $(B, 32, 768)$ from its final layer and spatially detailed features $(B, 256, 384)$ from its penultimate layer. Because of the compact spatial dimensions of DCFormer's outputs, our model can adopt the low-high hybrid structure without exceeding the token length limit of the LLM. This is crucial, as excessively long image token sequences can crowd out text tokens, leading to truncated questions or reports during inference. To leverage both types of features, we utilize two parallel MLP-Mixer modules—one for each layer. After passing through N Mixer layers, the resulting feature outputs are concatenated into a unified sequence of image tokens. In parallel, the input text is tokenized into a sequence of token IDs using the vocabulary of the LLM. This step follows the LLaVA [20] framework, which requires token embeddings for multimodal fusion. The resulting image and text tokens are then fused and passed to the LLM for joint reasoning.

This hybrid design enables the projector to fuse high-level abstract semantics with low-level spatial details, improving multi-modal alignment. Unlike prior approaches that downsample features at the expense of context, our method preserves diverse information without increasing token length. As a result, our projector produces richer and more semantically aligned joint representations, significantly enhancing radiology report generation and VQA. To the best of our knowledge, this work is the first to introduce a dual stream low-high hybrid MLP-Mixer for multimodal fusion in a 3D vision-language model. Unlike the original single-stream MLP-Mixer, our design processes both low-level spatial features and high-level semantic representations extracted from the 3D image encoder. This configuration enables the model to align textual inputs with both fine-grained anatomical details and high-level semantic context, which is critical for medical vision-language understanding. Empirically, it outperforms standard single-scale MLP-Mixer across multiple tasks.

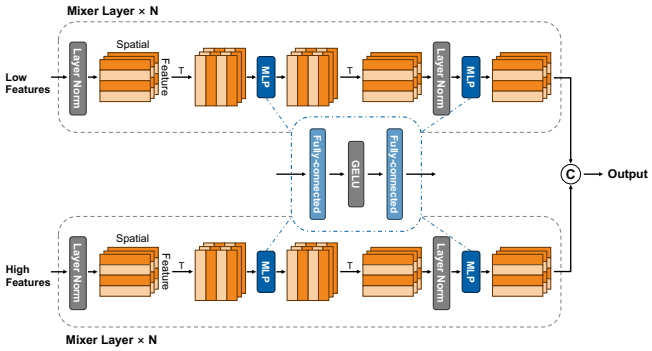


Figure 2: Multi-modal projector based on multi-scale MLP-Mixer.

We conducted this pretraining stage on both the M3D-Cap and M3D-VQA datasets. To mitigate binary-response bias and ensure richer semantic grounding, we excluded yes/no questions from M3D-VQA.

3.2.3 VLM Fine-tuning

In this stage, we fine-tune the multi-modal projector and the LLM for the report generation and VQA. Instead of fine-tuning the entire LLM, we adopt Low-Rank Adaptation (LoRA) [13], which significantly reduces the number of trainable parameters by introducing low-rank updates to the model's weights. This approach enables efficient adaptation while preserving the LLM's general knowledge.

Specifically, LoRA modifies a weight matrix $W_0 \in \mathbb{R}^{d \times k}$ by adding a learnable low-rank update:

$$W = W_0 + \Delta W, \quad \text{where} \quad \Delta W = BA \tag{9}$$

where $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$ are low-rank matrices with $r \ll \min(d, k)$. This decomposition allows task-specific tuning using far fewer parameters than full fine-tuning, resulting in faster training and lower memory usage.

During fine-tuning, only the weights of the multi-modal projector and the LoRA modules are updated, while all other parameters of the LLM remain frozen. The LLM

is initialized with Qwen2.5-7B-Instruct [42], a model pretrained for instruction-following tasks. By jointly optimizing the LoRA parameters and the multi-modal projector, our model effectively adapts to the characteristics of the M3D-Cap and M3D-VQA datasets.

3.3. Evaluation Metrics

We evaluate image-text retrieval performance using Recall@K (R@K) metrics, which quantify the percentage of correct matches within the top K retrieved items. Specifically, R@1 represents the proportion of queries where the correct match is ranked first, while R@5 and R@10 indicate the proportion of queries where the correct match appears within the top 5 or 10 results, respectively.

We evaluate the quality of radiology report generation and open-ended VQA using four key metrics: BLEU [25], ROUGE [17], METEOR [3], and BERTScore [46]. BLEU measures precision by evaluating the overlap of n-grams between the generated text and the reference text, focusing on how many words or phrases in the generated report are found in the reference text. The BLEU score is calculated as:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log p_n\right) \tag{10}$$

where BP (brevity penalty) is defined as:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \tag{11}$$

Here, c is the length of the generated text, r is the length of the reference text, w_n is the weight for n-grams, and p_n is the precision for n-grams of size n.

ROUGE emphasizes recall, assessing how much of the reference text appears in the generated output, which is particularly useful for summarization tasks to ensure key details are retained. The ROUGE score is calculated as:

$$\text{ROUGE} = \frac{\sum_{n=1}^{N} \text{Recall}_n}{\sum_{n=1}^{N} \text{Reference}_n} \tag{12}$$

where Recall_n is the recall of n-grams in the generated report, and Reference_n is the reference n-grams.

METEOR improves upon BLEU and ROUGE by considering synonyms and stemming, balancing precision and recall to provide a more flexible and comprehensive measure. The METEOR score is calculated as:

$$\text{METEOR} = \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

where Precision and Recall are calculated based on the number of matching words between the generated and reference reports.

BERTScore utilizes deep learning-based contextual embeddings to evaluate semantic similarity. The BERTScore is calculated as:

$$\text{BERT-Score} = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{j=1}^{M} \text{cosine}(e_i, e_j)}{M} \tag{14}$$

where e_i and e_j are the contextual embeddings of the generated and reference words, respectively, and M is the number of reference words.

3.4. Implementation Details

We implemented Med3DVLM using PyTorch [26] and Hugging Face Transformers [38], leveraging DeepSpeed ZeRO2 and BF16 precision on 8 NVIDIA A100 80GB GPUs. We used the AdamW optimizer [21] with a warmup ratio of 0.03 and a cosine learning rate scheduler.

3.4.1 Contrastive Pretraining

During contrastive pretraining on the M3D-Cap dataset, we used image-text pairs, where images were 3D CT volumes and texts were corresponding radiology reports. We computed similarity scores between pooled image and text embeddings to optimize a contrastive loss. We resized 3D CT volumes to $128 \times 256 \times 256$. We used DCFormer-small [1] as the vision encoder, and ClinicalBERT [34] as the text encoder. Image features were aggregated via mean pooling, while text features were achieved from the [CLS] token, as it captures the global semantic representation of the entire sequence in ClinicalBERT, making it suitable for contrastive alignment with image features. Both were projected to a shared 768-dimensional space. We used a batch size of 64, learning rate of 1×10^{-4}, weight decay of 0.1, and trained for 100 epochs.

3.4.2 VLM Pretraining

For VLM pretraining, we used image-question-answer triplets, where the image was 3D CT volume, the question was a natural language question query, and the answer was the corresponding free-text response. These triplets were sourced from the M3D-Cap and M3D-VQA datasets (excluding yes/no questions) dataset. The DCFormer-small encoder was kept frozen. To align the image encoder and the LLM, we introduced a hybrid MLP-Mixer projector that fused low-level features from the penultimate encoder layer (256×384) and high-level features from the final layer (32×768) of the image encoder. Each feature stream passed through N Mixer layers, and the resulting tokens were concatenated and paired with text tokens as input to the LLM. This stage used a batch size of 16, learning rate of 1×10^{-4}, no weight decay, and was trained for 3 epochs.

3.4.3 VLM Fine-tuning

For VLM fine-tuning, We continued training using image–question–answer triplets from M3D-Cap and M3D-VQA, now including yes/no questions. The image encoder and MLP-Mixer projector were reused. The model was fine-tuned to generate answers conditioned on the image and question. We used LoRA [13] with a rank of 16, scaling factor $\alpha = 32$, and dropout of 0.05, updating only the LoRA modules and the projector while keeping all other weights frozen. We fine-tuned the model using a batch size of 8, learning rate of 5×10^{-5}, no weight decay, and training for 5 epochs. All M3D-LaMed results are reported as in the original paper [2].

3.4.4 Inference and Evaluation

During inference, we evaluated two tasks: report generation and visual question answering. In both tasks, the input consisted of a 3D CT volume and a natural language prompt. For report generation, the prompt was a fixed instruction such as "Please provide a radiology report for the given CT volume." The model generated a free-text radiology report (see Fig 3 for an example). In VQA, the prompt was a question about the image. In open-ended VQA, the model generated free-form answers. In close-ended VQA, multiple answer options were included in the prompt, and the model selected the most appropriate ones (see Fig 4 for examples).

4. Results

4.1. Image-Text Retrieval

Med3DVLM significantly outperforms M3D-LaMed in image-to-text and text-to-image retrieval tasks, as shown in Table 1. It achieves the highest Recall@1 across all test set sizes (100, 500, 1000, and 2000). These consistent gains highlight its ability to learn semantically rich representations from high-resolution volumetric data and align them with corresponding text descriptions.

Notably, the performance gap between Med3DVLM and M3D-LaMed widens with larger test sets, suggesting that Med3DVLM generalizes better under more challenging and diverse retrieval scenarios. This trend indicates robustness not only to variation in anatomical content and imaging conditions but also to increasing data scale—a critical requirement for deployment in real-world clinical settings. In such environments, retrieval systems must support large-scale databases while maintaining high precision.

4.2. Radiology Report Generation

4.2.1 Quantitative Performance Evaluation

Med3DVLM achieves state-of-the-art performance in radiology report generation, as shown in Table 2. It outperforms all baselines across BLEU, ROUGE, and METEOR, with particularly notable gains in METEOR, which captures both fluency and content relevance. These results suggest that Med3DVLM can generate clinically coherent, well-structured descriptions of complex imaging findings.

Interestingly, although Med3DVLM and M3D-LaMed achieve similar BERTScores (88.11 vs. 88.46), BERTScore primarily reflects abstract similarity in latent space and does not fully capture linguistic precision or clinical completeness. In contrast, Med3DVLM shows substantial gains on semantically grounded metrics, with METEOR improving from 14.38% to 36.42% and ROUGE from 19.55% to 40.25%. These improvements indicate that Med3DVLM generates reports with more accurate lexical choices, clearer structure, and better alignment with clinical content. The discrepancy between BERTScore and traditional n-gram metrics highlights the importance of evaluating both semantic fidelity and textual quality, which are essential for accurate and readable medical reporting.

4.2.2 Qualitative Analysis of Generated Reports

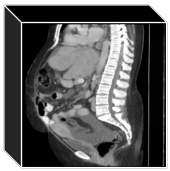
Figure 3 compares the radiology reports generated by Med3DVLM and M3D-LaMed on a chest CT. Med3DVLM correctly identifies key abnormalities, including multifocal hepatic mass lesions, portal vein thrombosis, and occlusive filling defects, aligning well with the ground truth. It also captures scattered calcifications and small-volume pelvic free fluid, findings not explicitly mentioned but plausibly related to the documented pathology. In contrast, M3D-LaMed fails to detect any liver pathology, underscoring its limitations in aligning imaging features with diagnostic text.

Despite its improvements, Med3DVLM generates incorrect findings, such as hepatic steatosis, contrast enhancement, and iliac arteriovenous malformation, suggesting a tendency to overgeneralize common radiological patterns. Mentions of left ureter abnormalities and renal cortical thinning appear to be hallucinated, indicating the need for better factual grounding. M3D-LaMed, while not generating outright hallucinations, produces findings entirely unrelated to liver pathology, such as bilateral ovarian vein enlargement and para-uterine venous varicosities, suggesting weaker cross-modal alignment.

Table 1: Image-text retrieval performance on the M3D dataset. R@K: Recall@Top K. IR: image retrieval. TR: text retrieval. DCFormer-S: small version of DCFormer.

Methods		PMC-CLIP				M3D-LaMed				DCFormer-S SigLIP (Ours)			
Test Samples		100	500	1000	2000	100	500	1000	2000	100	500	1000	2000
IR	R@1	9.00	4.40	1.90	1.15	64.00	39.60	27.30	19.10	**92.00**	**77.20**	**69.30**	**61.00**
	R@5	28.00	12.80	7.60	4.35	**95.00**	76.20	61.10	47.45	**95.00**	**92.80**	**91.40**	**87.20**
	R@10	45.00	18.80	12.10	7.60	**99.00**	87.20	76.10	62.25	96.00	**95.20**	**93.90**	**91.25**
TR	R@1	18.00	7.60	4.60	3.15	70.00	40.40	26.60	18.45	**90.00**	**78.40**	**71.30**	**63.90**
	R@5	47.00	20.20	13.00	8.55	**95.00**	74.20	61.80	47.30	**95.00**	**93.40**	**91.40**	**88.05**
	R@10	59.00	31.00	19.80	13.55	**98.00**	87.00	75.30	62.15	95.00	**95.60**	**94.10**	**91.90**

Report Generation



<Ground Truth>: Liver shows the following: cirrhotic changes with surface nodularity and hypertrophied caudate lobe multiple bi-lobar patches of low attenuation (<10 HU attenuation in the non contrast phase and <25 HU in the portal venous phase), suggestive of hepatic steatosis right hepatic lobe segment VI and left hepatic lobe segment IV a patches of contrast enhancement showing contrast enhancement at the early enhancement at the arterial and portal phases with no contrast wash out at the delayed phase, likely inflammatory in nature signs of fibrosis evident by surface nodularity, widened preportal and gallbladder fossae Portal hypertension and splenomegaly. Mild abdominal and pelvic ascites.

<M3D-LaMed>: Multiple venous varicosities were seen in bilateral para uterine/parametrial regions, more dilated on the left side. The largest diameter of the left para-uterine vein measures 1.0cm. Bilateral ovarian veins are enlarged, measuring up to 0.9cm. There are duplications of bilateral ovarian veins which are the common normal variants. Left ovarian vein drains into left renal vein. Right ovarian vein drains into inferior vena cava. Reduced aortic-SMA angle where significant compression onto left renal vein between superior mesenteric artery (SMA) and aorta (nutcracker phenomenon) -mid part of left renal vein.

<Med3DVLM>: **Multifocal hepatic mass lesions seen involving the right and left lobes of the liver. Scattered calcifications within the lesions.** No evidence of intrahepatic gas. **Large right and left portal vein thrombosis. Occlusive filling defects seen within the left and right main portal veins.** Multiple right and left iliac lymph nodes. Left iliac arteriovenous malformation. Left atretic ureter with a polypoidal bladder diverticulum seen at the left VUJ. Left hydroureter and hydronephrosis. Left renal cortical thinning. No left renal vein opacification. Occluded left proximal common iliac vein. Occlusion of the left external iliac artery with a left peroneal flap artery. **Small volume pelvic free fluid.** Left scrotal edema.

Figure 3: An example of generated radiology reports by Med3DVLM and M3D-LaMed. Med3DVLM demonstrates improved alignment with clinical findings, while M3D-LaMed generates irrelevant or incorrect content.

Table 2: Report generation performance on the M3D-Cap dataset.

Method	BLEU	ROUGE	METEOR	BERTScore
RadFM	12.23	16.49	11.57	87.93
M3D-LaMed (Linear)	14.49	19.25	14.11	88.32
M3D-LaMed (MLP)	15.15	19.55	14.38	**88.46**
Med3DVLM (Ours)	**36.88**	**40.25**	**36.42**	88.11

4.3. Visual Question Answering

4.3.1 Quantitative Performance Evaluation

Med3DVLM achieves state-of-the-art performance in both open-ended and close-ended VQA tasks, as shown in Tables 3 and 4. The model demonstrates particularly strong results in clinically important categories such as organ and abnormality identification, highlighting its ability to capture fine-grained anatomical and contextual cues in 3D volumes.

In open-ended VQA, Med3DVLM improves the METEOR score from 33.58% to 36.76% (a 9.5% improvement) and the ROUGE score from 52.39% to 56.31% (a 7.5% improvement) compared to M3D-LaMed. These metrics offer a more semantically grounded evaluation than simple lexi-

cal overlap, capturing both the accuracy and fluency of generated responses, which are essential for answering complex clinical queries with clarity and relevance. The results suggest that Med3DVLM produces free-text responses that are not only better aligned with the ground truth but also more structured and clinically coherent.

In close-ended VQA, Med3DVLM achieves 79.75% accuracy, surpassing M3D-LaMed's 75.78% by 5.2%. Although the margin is smaller than in open-ended VQA, the improvement remains clinically meaningful: even modest gains in accuracy can reduce diagnostic errors and enhance the reliability of automated decision support systems in clinical practice.

4.3.2 Qualitative Analysis of VQA Accuracy

Figure 4 presents a qualitative comparison of Med3DVLM and M3D-LaMed in open-ended and closed-ended VQA. In open-ended VQA, both M3D-LaMed and Med3DVLM correctly recognizes the axial imaging plane and localizes a mass lesion in the liver. However, in more complex cases, such as lesion localization in the cranial fossa, Med3DVLM provides a partially correct answer, whereas M3D-LaMed misidentifies the region entirely. While Med3DVLM significantly improves answer accuracy, errors in lesion localization indicate room for enhancement in fine-grained spatial reasoning and uncertainty calibration.

In closed-ended VQA, Med3DVLM consistently outperforms M3D-LaMed by correctly identifying the pleural effusion in the right lung, the mass lesion in the stomach, and the portal venous phase of the CT scan. In contrast, M3D-LaMed misclassifies the pleural effusion as being in the left lung, the stomach lesion as adenocarcinoma, and the CT phase as contrast phase. These errors indicate M3D-LaMed's limitations in recognizing imaging features and aligning them with clinical knowledge.

4.4. Ablation Study

4.4.1 Effect of Core Components

We first analyze the impact of the three core components of Med3DVLM: the DCFormer vision encoder, the SigLIP contrastive loss, and the multi-scale MLP-Mixer projector. The results are summarized in Table 5, Table 6, and Table 7, corresponding to radiology report generation, open-ended VQA, and closed-ended VQA, respectively.

In the radiology report generation task, the full model achieves a METEOR score of 36.42%. Removing the MLP-Mixer leads to a drop in performance to 23.78%, while removing the SigLIP loss results in a lower METEOR score of 13.48%. When the DCFormer encoder is removed, the MLP-Mixer must also be excluded since it depends on the multi-scale feature outputs from DCFormer. In this case, the METEOR score degrades severely, reaching only 8.88%. Similar trends are observed in BLEU, ROUGE, and BERTScore, highlighting the critical role of all three components in producing clinically accurate and semantically aligned text.

For open-ended VQA, removing both DCFormer and MLP-Mixer leads to a drastic drop in performance, reducing the mean METEOR score from 36.76% to just 1.38%. In contrast, removing only the MLP-Mixer causes a modest drop to 34.65%, while removing SigLIP results in a minimal decrease to 36.36%. Similarly, for closed-ended VQA, the full model achieves a mean accuracy of 79.75%. Removing SigLIP has little effect (79.71%), and removing the MLP-Mixer causes a small drop to 78.58%. In contrast, removing both DCFormer and MLP-Mixer results in a sharp decline to 1.26% accuracy. These results suggest that while all three components contribute to overall VQA performance, the DCFormer encoder is essential for reliable visual understanding, and the MLP-Mixer plays an important role in multi-modal reasoning.

In summary, each core component of Med3DVLM contributes to overall performance, with varying importance across tasks. The DCFormer encoder is essential for robust 3D representation across all settings, while the MLP-Mixer plays a key role in multi-modal reasoning, particularly for generative tasks. The SigLIP loss enhances alignment stability, especially in report generation.

4.4.2 Impact of Vision Encoder and Contrastive Loss

We evaluated the impact of different vision encoders (ViT3D vs. DCFormer-S) and contrastive loss functions (CLIP vs. SigLIP) on image-text retrieval. As shown in Table 8, both image retrieval (IR) and text retrieval (TR) benefit significantly from using DCFormer-S over ViT3D, particularly at larger test sizes. For example, with 2,000 test samples and SigLIP loss, DCFormer-S achieves an R@1 of 61.00% for IR and 63.90% for TR, compared to 38.65% and 39.70% respectively with ViT3D.

Across all encoder types and test sizes, switching from CLIP to SigLIP consistently improves performance. For example, with the ViT3D encoder and 2,000 test samples, SigLIP improves IR R@1 from 19.10% to 38.65%, and TR R@1 from 19.40% to 39.70%. With the DCFormer-S encoder, the gains are also evident: IR R@1 increases from 29.85% to 61.00%, and TR R@1 from 32.60% to 63.90%. These improvements are especially pronounced at larger evaluation scales, where SigLIP enhances alignment stability without requiring large negative batches.

4.4.3 Impact of Multi-Modal Projectors

We evaluated the impact of different multi-modal projector designs on radiology report generation, open-ended VQA, and closed-ended VQA. The 2×MLP variant consists of a

Table 3: Open-ended visual question answering performance on the M3D-VQA dataset.

Open-ended VQA							
Method	Metric	Plane	Phase	Organ	Abnormality	Location	Mean
RadFM	BLEU	14.24	14.25	14.24	15.64	23.58	16.39
	ROUGE	25.40	25.41	25.38	25.38	29.09	26.13
	METEOR	20.62	20.63	20.61	20.60	24.19	21.33
	BERTScore	92.68	92.04	86.79	85.84	86.26	88.72
M3D-LaMed	BLEU	98.37	74.41	34.20	15.91	24.00	49.38
	ROUGE	98.42	78.63	37.87	19.27	27.74	52.39
	METEOR	49.20	63.58	23.78	12.83	18.50	33.58
	BERTScore	99.47	95.55	88.97	86.08	87.60	91.53
Med3DVLM (Ours)	BLEU	**98.85**	**78.17**	**40.22**	**18.99**	**25.66**	**52.38**
	ROUGE	**98.89**	**84.20**	**45.22**	**23.27**	**29.99**	**56.31**
	METEOR	**49.43**	**68.50**	**29.32**	**16.21**	**20.32**	**36.76**
	BERTScore	**99.83**	**96.47**	**90.47**	**86.27**	**87.88**	**92.18**

Table 4: Comparison of closed-ended visual question answering performance with state-of-the-art methods on the M3D-VQA dataset. The results are reported in terms of accuracy.

Close-ended VQA						
Methods	Plane	Phase	Organ	Abnormality	Location	Mean
RadFM	19.65	28.70	16.80	18.92	14.88	19.79
M3D-LaMed	98.80	79.75	74.75	66.65	58.94	75.78
Med3DVLM (Ours)	**99.15**	**87.50**	**77.45**	**70.17**	**64.49**	**79.75**

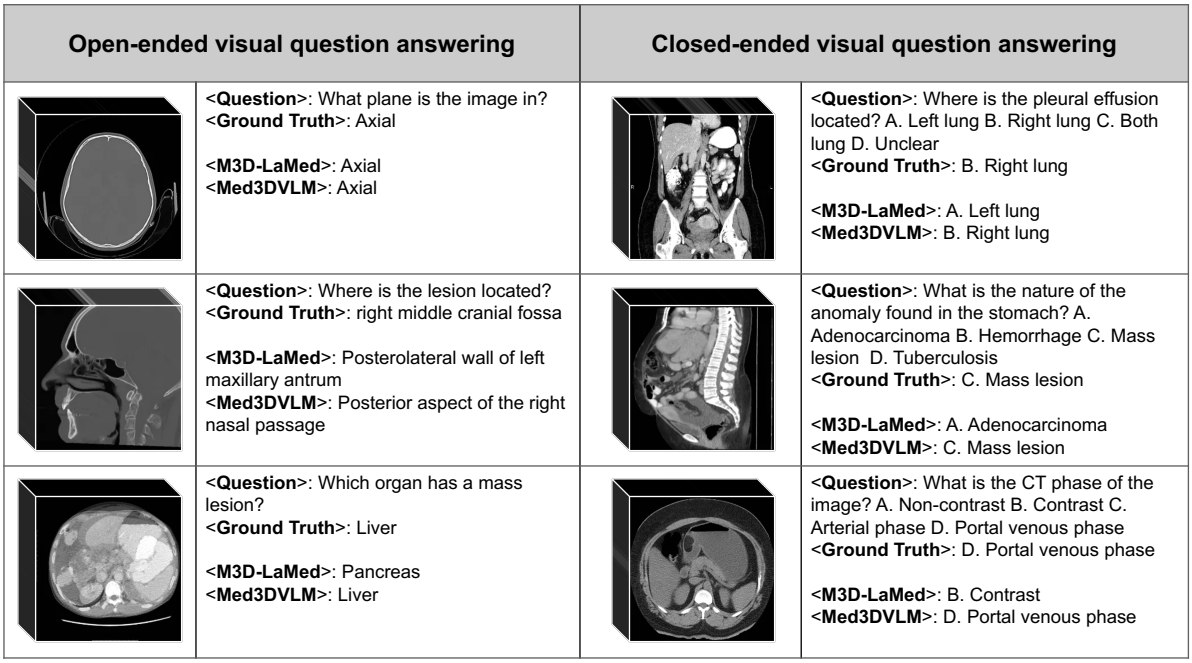


Figure 4: Examples of open-ended and closed-ended VQA results by Med3DVLM and M3D-LaMed.

Table 5: Effect of core Med3DVLM components on radiology report generation.

DCFormer	SigLIP Loss	MLP-Mixer	BLEU	ROUGE	METEOR	BERT
✓		✓	24.47	28.80	23.78	85.75
✓	✓		14.31	18.14	13.48	83.86
	✓		8.62	11.23	8.88	77.07
✓	✓	✓	**36.88**	**40.25**	**36.42**	**88.11**

Table 6: Effect of core Med3DVLM components on open-ended VQA performance.

DCFormer	SigLIP Loss	MLP-Mixer	Metric	Plane	Phase	Organ	Abnormality	Location	Mean
✓		✓	BLEU	**99.00**	77.28	39.13	**19.73**	24.85	52.00
			ROUGE	**99.05**	**84.23**	43.48	**23.85**	29.48	56.02
			METEOR	**49.52**	68.31	27.23	**16.69**	20.07	36.36
			BERT	**99.85**	96.30	90.33	**86.29**	87.74	92.10
✓	✓		BLEU	98.60	75.00	34.34	17.76	24.55	50.05
			ROUGE	98.67	81.55	38.59	21.45	28.84	53.82
			METEOR	49.37	65.18	24.28	14.72	19.69	34.65
			BERT	99.79	95.87	89.44	85.93	87.71	91.75
	✓		BLEU	0.00	0.17	0.19	0.73	0.57	0.33
			ROUGE	0.08	0.43	0.96	1.54	1.46	0.89
			METEOR	0.07	0.90	1.62	2.19	2.14	1.38
			BERT	73.18	74.68	74.47	74.01	74.24	74.12
✓	✓	✓	BLEU	98.85	**78.17**	**40.22**	18.99	**25.66**	**52.38**
			ROUGE	98.89	84.20	**45.22**	23.27	**29.99**	**56.31**
			METEOR	49.43	**68.50**	**29.32**	16.21	**20.32**	**36.76**
			BERT	99.83	**96.47**	**90.47**	86.27	**87.88**	**92.18**

Table 7: Effect of core Med3DVLM components on closed-ended VQA accuracy.

DCFormer	SigLIP	MLP-Mixer	Plane	Phase	Organ	Abnormality	Location	Mean
✓		✓	**99.30**	**87.95**	77.40	69.90	63.99	79.71
✓	✓		98.85	84.95	74.80	69.82	64.46	78.58
	✓		1.60	1.05	1.35	1.03	1.28	1.26
✓	✓	✓	99.15	87.50	**77.45**	**70.17**	**64.49**	**79.75**

simple two-layer multilayer perceptron applied to global visual features, using a linear activation followed by GELU and another linear transformation. The 2×MLP-H variant incorporates hierarchical features from the final and penultimate layers of the vision encoder; each feature map is independently processed by a 2×MLP block, and their outputs are concatenated. The 1×MLP-Mixer-H design replaces the MLPs with a single MLP-Mixer block that jointly processes the concatenated hierarchical features, enabling token and channel mixing. The 2×MLP-Mixer-H configuration, used in our final model, stacks two MLP-Mixer blocks and includes an additional hidden projection layer to further enhance multi-scale feature interaction. To provide a complete view of computational efficiency, we also report the number of parameters and FLOPs for each projector variant in Table 9

As shown in Table 9, the 2×MLP-Mixer-H configuration achieves the highest METEOR score of 36.42% on report generation, significantly outperforming simpler alternatives such as 2×MLP (15.10%). The single-scale 1×MLP-Mixer-H achieves a much higher METEOR of 23.25% compared to MLP-based projectors, although it still lags behind the proposed multi-scale design. These results highlight the importance of multi-scale feature integration for generating coherent and semantically aligned clinical reports.

For both open-ended and closed-ended VQA, the 2×MLP-Mixer-H projector achieves the best overall performance (Tables 10 and 11). In open-ended VQA, it improves the mean METEOR from 34.39% (2×MLP) to 36.76%, with strong gains in semantically challeng-

Table 8: The impact of different vision encoders and loss functions on image-text retrieval. DCFormer-S represents small version of DCFormer.

Methods		ViT3D CLIP				ViT3D SigLIP			
Test Samples		100	500	1000	2000	100	500	1000	2000
IR	R@1	63.00	33.40	25.90	19.10	75.00	55.40	46.20	38.65
	R@5	89.00	67.80	57.20	44.70	91.00	84.40	79.50	70.50
	R@10	95.00	81.00	71.60	59.35	94.00	89.80	87.10	79.45
TR	R@1	57.00	32.00	25.50	19.40	78.00	59.60	49.40	39.70
	R@5	90.00	67.40	56.90	45.35	91.00	84.20	77.40	69.20
	R@10	94.00	81.80	70.70	59.10	93.00	88.80	85.50	78.40
Methods		DCFormer-S CLIP				DCFormer-S SigLIP			
Test Samples		100	500	1000	2000	100	500	1000	2000
IR	R@1	76.00	51.80	40.80	29.85	**92.00**	**77.20**	**69.30**	**61.00**
	R@5	94.00	86.00	74.60	64.60	**95.00**	**92.80**	**91.40**	**87.20**
	R@10	**97.00**	91.40	84.90	75.95	96.00	**95.20**	**93.90**	**91.25**
TR	R@1	83.00	52.60	43.50	32.60	**90.00**	**78.40**	**71.30**	**63.90**
	R@5	94.00	86.00	75.00	64.10	**95.00**	**93.40**	**91.40**	**88.05**
	R@10	**98.00**	91.60	85.70	76.70	95.00	**95.60**	**94.10**	**91.90**

Table 9: The impact of multi-modal projectors on radiology report generation. All models use DCFormer-small and the Qwen-2.5-7B-Instruct LLM. H: low high hybrid.

Method	BLEU	ROUGE	METEOR	BERTScore	Parameters	Flops
2xMLP	15.63	19.86	15.10	84.24	15.60 M	0.50 G
2xMLP-H	15.99	20.25	15.68	84.31	16.98 M	4.14 G
1xMLP-Mixer-H	23.93	27.70	23.25	85.73	29.91 M	3.85 G
2xMLP-Mixer-H	**36.88**	**40.25**	**36.42**	**88.11**	47.22 M	6.14 G

ing categories like "organ" and "abnormality." In closed-ended VQA, although the performance gap is narrower, the 2×MLP-Mixer-H still achieves the highest mean accuracy of 79.75%, compared to 78.65% with 2×MLP. These results suggest that while classification tasks rely more on pattern recognition, generative tasks benefit significantly from multi-scale feature fusion, and the MLP-Mixer projector enhances semantic reasoning and alignment across both task types.

5. Discussion

5.1. Addressing Limitations of Existing Models

While previous 3D VLMs such as M3D-LaMed and RadFM have made significant advances, they remain constrained by computational inefficiency, limited capacity for fine-grained spatial reasoning, and suboptimal multi-modal alignment. Med3DVLM addresses these gaps through three targeted innovations. First, Med3DVLM uses DCFormer to decompose 3D convolutions to reduce complexity while preserving spatial detail, allowing for the processing of high-resolution image volumes. Second, SigLIP improves contrastive learning on small, semantically dense medical datasets by avoiding reliance on large negative batches. Third, the dual-stream MLP-Mixer projector effectively fuses low- and high-level image features with text embeddings, enriching the semantic alignment with the LLM.

5.2. Clinical Implications

The integration of Med3DVLM into clinical workflows offers strong potential to improve medical imaging interpretation, decision support, and patient care. Its robust performance in image-text retrieval enables accurate alignment between 3D imaging studies and corresponding textual reports, facilitating rapid access to similar prior cases. This capability supports comparative analysis, longitudinal tracking, and second-opinion workflows, and can enhance diagnostic accuracy and consistency. In educational settings, it may also serve as a valuable tool for case-based learning and clinical training. Med3DVLM also generates accurate, fluent radiology reports directly from 3D volumetric data, reducing documentation burden—especially in

Table 10: The impact of multi-modal projectors on open-ended visual question answering. All models used the DCFormer-small vision encoder and Qwen 2.5-7B-Instruct model. H: low high hybrid.

Open-ended VQA							
Method	Metric	Plane	Phase	Organ	Abnormality	Location	Mean
2xMLP	BLEU	98.67	74.23	33.95	16.92	23.79	49.51
	ROUGE	98.72	80.72	38.46	20.97	28.04	53.38
	METEOR	49.35	64.69	24.30	14.63	18.97	34.39
	BERTScore	99.80	95.75	89.50	85.88	87.64	91.71
2xMLP-H	BLEU	**98.86**	74.46	34.37	16.96	24.49	49.83
	ROUGE	98.91	80.87	38.55	20.82	28.81	53.59
	METEOR	49.48	64.88	24.42	14.36	19.29	34.49
	BERTScore	**99.83**	95.75	89.50	85.88	87.80	91.75
1xMLP-Mixer-H	BLEU	98.85	76.37	38.87	17.47	24.62	51.24
	ROUGE	**98.93**	82.69	42.93	21.46	28.94	54.99
	METEOR	**49.51**	67.00	26.86	15.19	19.63	35.64
	BERTScore	**99.83**	96.16	90.29	85.99	87.80	92.01
2xMLP-Mixer-H	BLEU	98.85	**78.17**	**40.22**	**18.99**	**25.66**	**52.38**
	ROUGE	98.89	**84.20**	**45.22**	**23.27**	**29.99**	**56.31**
	METEOR	49.43	**68.50**	**29.32**	**16.21**	**20.32**	**36.76**
	BERTScore	**99.83**	**96.47**	**90.47**	**86.27**	**87.88**	**92.18**

Table 11: The impact of multi-modal projectors on closed-ended visual question answering. All models used the DCFormer-small vision encoder and Qwen 2.5-7B-Instruct LLM. H: low high hybrid.

Close-ended VQA						
Methods	Plane	Phase	Organ	Abnormality	Location	Mean
2xMLP	98.75	84.35	75.90	69.55	64.70	78.65
2xMLP-H	98.75	86.40	75.05	69.90	62.67	78.55
1xMLP-Mixer-H	**99.30**	86.10	77.05	**70.19**	63.92	79.31
2xMLP-Mixer-H	99.15	**87.50**	**77.45**	70.17	**64.49**	**79.75**

high-volume clinical environments. Its substantial gains in METEOR and ROUGE scores indicate strong alignment with expert-level interpretations, making it a practical tool for assisting in preliminary report drafting and standardizing report quality. Finally, the model's VQA functionality allows clinicians to interact with 3D scans using natural language. This enables task-specific, context-aware responses that support anatomical identification, diagnostic clarification, and real-time clinical decision-making—ultimately reducing the risk of oversight and enhancing confidence in image interpretation.

5.3. Challenges and Future Directions

While Med3DVLM sets a new standard in 3D VLM, several challenges remain. The model occasionally generates hallucinated content in radiology reports, such as references to unrelated anatomical findings, highlighting the persistent issue of factual grounding in generative VLMs. Additionally, although performance in open-ended VQA improved markedly, some errors in lesion localization and answer precision persist, suggesting the need for enhanced spatial reasoning and more robust alignment between visual features and textual outputs.

While the M3D dataset provides a large-scale and diverse benchmark, it primarily consists of CT images and focuses on a limited set of tasks including retrieval, report generation, and VQA. To ensure broader generalization and real-world applicability, future evaluation should incorporate other imaging modalities such as MRI and ultrasound, as well as datasets that differ in patient demographics, clinical settings, and scanner protocols. In addition, extending the scope of evaluation to include tasks such as medical image synthesis, modality translation, and segmentation would offer a more comprehensive assessment of the model's robustness and clinical utility across a wider range of scenarios.

As for future directions, incorporating structured clinical knowledge could help reduce hallucinations and improve interpretability. Furthermore, as noted above, validating Med3DVLM across diverse clinical datasets and imaging protocols will be critical to ensuring its reliability in real-world settings. Expanding the model to support uncertainty-aware response generation may also enhance its trustworthiness in clinical decision support.

5.4. Computational Complexity

We compare Med3DVLM and M3D-LaMed in terms of parameter count and FLOPs, as summarized in Table 12. Med3DVLM replaces the 3D Vision Transformer in M3D-LaMed with the DCFormer encoder, reducing the vision backbone from 87.4M parameters and 253.23G FLOPs to 18.2M parameters and 21.59G FLOPs. Although Med3DVLM uses a larger multi-modal projector (47.2M vs. 19.9M parameters), this overhead is minor compared to the efficiency gains from the encoder. The total model size increases slightly (7.6B vs. 6.9B) due to a larger LLM, which is fine-tuned efficiently using LoRA.

We do not report direct comparisons of inference time and memory usage, as M3D-LaMed includes an additional segmentation module not present in our model. This makes runtime comparisons non-equivalent. To ensure fairness, we focus on theoretical complexity metrics that isolate the vision-language components.

Table 12: Comparison of parameters and FLOPs for each module in Med3DVLM. M3D-LaMed values are reported from the original paper.

Module	M3D-LaMed		Med3DVLM	
	Params	FLOPs	Params	FLOPs
3D Image Encoder	87.4M	253.23G	18.2M	21.59G
Multimodal Projector	19.9M	5.10G	47.2M	6.14G
LLM with LoRA	6.7B	-	7.6B	-
All	6.9B	-	7.6B	-

6. Conclusion

We presented Med3DVLM, a vision-language model designed specifically for 3D medical image analysis. Med3DVLM integrates an efficient volumetric encoder (DCFormer), a sigmoid-based contrastive learning strategy (SigLIP), and a dual-stream MLP-Mixer projector. Extensive experiments on the M3D dataset demonstrate that Med3DVLM achieves state-of-the-art performance in image-text retrieval, radiology report generation, and visual question answering. Notably, Med3DVLM maintains high accuracy while remaining computationally efficient, making it suitable for deployment in real-world clinical workflows. These results underscore the promise of 3D VLMs as foundational tools for building automated, interpretable, and multimodal medical AI systems.

References

[1] Gorkem Can Ates, Kuang Gong, and Wei Shao. Dcformer: Efficient 3d vision-language modeling with decomposed convolutions. *arXiv preprint arXiv:2502.05091*, 2025. 2, 3, 4, 6

[2] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024. 1, 2, 3, 5, 7

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[4] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022. 2

[5] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 5

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 4

[7] Wenjie Dong, Shuhao Shen, Yuqiang Han, Tao Tan, Jian Wu, and Hongxia Xu. Generative models in medical visual question answering: A survey. *Applied Sciences*, 15(6):2983, 2025. 3

[8] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021. 1

[9] Deepak Gupta, Swati Suman, and Asif Ekbal. Hierarchical deep multi-modal network for medical visual question answering. *Expert Systems with Applications*, 164:113993, 2021. 3

[10] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Omer Faruk Durugol, Bastian Wittmann, Tamaz Amiranashvili, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024. 3

[11] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–486. Springer, 2024. 3

[12] Philipp Harzig, Yan-Ying Chen, Francine Chen, and Rainer Lienhart. Addressing data bias problems for chest x-ray image report generation. *arXiv preprint arXiv:1908.02123*, 2019. 2

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5, 7

[14] Zhongzhen Huang, Yankai Jiang, Rongzhao Zhang, Shaoting Zhang, and Xiaofan Zhang. Cat: Coordinating anatomical-textual prompts for multi-organ and tumor segmentation. *arXiv preprint arXiv:2406.07085*, 2024. 2

[15] Haoran Lai, Zihang Jiang, Qingsong Yao, Rongsheng Wang, Zhiyang He, Xiaodong Tao, Wei Wei, Weifu Lv, and S Kevin Zhou. E3d-gpt: Enhanced 3d visual foundation for medical vision-language model. *arXiv preprint arXiv:2410.14200*, 2024. 2

[16] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 3

[17] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[18] 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. 1

[19] Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 210–220. Springer, 2021. 3

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 3, 4, 5

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[22] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024. 1

[23] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 3

[24] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 522–530. Springer, 2019. 3

[25] 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. 6

[26] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 6

[27] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023. 2

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3

[29] Fuji Ren and Yangyang Zhou. Cgmvqa: A new classification and generative model for medical visual question answering. *IEEE Access*, 8:50626–50636, 2020. 3

[30] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 2, 5

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[32] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024. 3

[33] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[34] Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642, 2023. 4, 6

[35] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023. 3

[36] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022. 2

[37] Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. A self-boosting framework for automated radiographic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2433–2442, 2021. 2

[38] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 6

[39] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*, 2023. 1, 2

[40] Yu Xin, Gorkem Can Ates, and Wei Shao. Text3dsam: Text-guided 3d medical image segmentation using sam-inspired architecture. In *CVPR 2025: Foundation Models for 3D Biomedical Image Segmentation*. 2

[41] Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 457–466. Springer, 2018. 2

[42] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 6

[43] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 4

[44] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 721–729. Springer, 2019. 2

[45] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2, 4

[46] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. 6

[47] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pages 2–25. PMLR, 2022. 2

[48] Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. One model to rule them all: Towards universal segmentation for medical images with text prompts. *arXiv preprint arXiv:2312.17183*, 2023. 2

[49] Wenbo Zheng, Lan Yan, Fei-Yue Wang, and Chao Gou. Learning from the guidance: Knowledge embedded meta-learning for medical visual question answering. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27*, pages 194–202. Springer, 2020. 3