



OPEN A data-efficient 3D medical vision-language model using only a 2D encoder

Yi Lian¹, Yanchun Xie¹✉, Yang Jiang², Lei Wang^{3,4} & Hailong Yu¹✉

The demonstrated success of Vision-Language Models in 2D medical image analysis has motivated the extension of their capabilities to 3D volumetric data for tasks such as report generation and visual question answering. A primary obstacle to this advancement is the reliance of current approaches on specialized 3D vision encoders, whose performance is constrained by the scarcity of large-scale annotated datasets. This paper presents a data-efficient framework that bypasses the need for a 3D encoder, instead leveraging a pre-trained 2D vision encoder to process volumetric data. Our pipeline sequentially refines the visual representation. First, a cosine similarity strategy prunes redundant 2D slices to improve computational efficiency. Next, a spatial-frequency fusion module integrates spatial and frequency-domain information to model inter-slice correlations from the 2D features. Finally, a fine-grained feature injection mechanism mitigates information loss during feature compression by re-introducing high-resolution details into the final visual tokens for the Large Language Model. Evaluated on public 3D benchmarks, our framework demonstrates superior performance, achieving a METEOR score of 50.13 on M3D-Cap report generation and 82.90% accuracy on M3D-VQA, significantly outperforming previous models. Our work demonstrates a scalable and efficient paradigm for 3D medical vision-language tasks that avoids the need for 3D-specific pre-training, offering a data-efficient alternative to data-intensive 3D encoders.

Keywords Medical visual-language models, 3D medical imaging, Data-efficient learning, 2D vision encoder

Medical Visual-Language Models (Med-VLMs)¹ are a class of artificial intelligence models designed to jointly analyze medical images and text. Their core workflow involves receiving a medical image and a text prompt as input and generating a corresponding textual response as output, enabling clinical support functions such as Medical Visual Question Answering (VQA)² and automatic radiology report generation³. Although Med-VLMs have achieved significant progress in 2D image analysis, clinical diagnosis increasingly relies on 3D images that provide complete spatial structures. Therefore, extending the capabilities of Med-VLMs from two to three dimensions is a critical step in advancing AI-assisted diagnostics.

However, the prevailing paradigm for 3D Med-VLMs^{4–6} generally relies on specialized 3D vision encoders⁷. These encoders typically require large-scale, meticulously annotated 3D medical datasets for pre-training, but the scarcity and high cost of such data limit the development, generalization, and practical clinical deployment of advanced 3D Med-VLMs. Overcoming this challenge of data scarcity has profound clinical implications. The heavy reliance on expensive and inaccessible 3D annotated datasets creates a barrier to deployment, particularly in resource-limited settings where acquiring high-quality volumetric data and expert annotations is often unfeasible. By reducing the dependency on massive 3D datasets, data-efficient approaches can pave the way for scalable applications in clinical environments with limited data availability. This shift is essential for bridging the digital divide and improving the real-world accessibility of AI-assisted diagnostics to broader populations. Furthermore, even with 3D encoders, existing methods exhibit common technical limitations. On one hand, they struggle to effectively manage the visual feature redundancy arising from dense slices in 3D images. On the other hand, their feature modeling primarily focuses on the spatial domain, leading to an insufficient ability to perceive critical diagnostic information such as intrinsic texture differences and structural heterogeneity within lesions⁸.

To address these limitations, we propose a data-efficient framework that leverages a pre-trained 2D vision encoder^{9,10} to perform 3D medical vision-language tasks. The framework begins by extracting slice-wise visual

¹Department of Orthopaedics, General Hospital of Northern Theater Command, Shenyang, China. ²Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China. ³CCTEG (Liaoning) Embodied Intelligence Technology Co., Ltd., Shenyang, China. ⁴CCTEG Robot Technology Co., Ltd., Shenzhen, China. ✉email: yanchunxie@foxmail.com; hailong7@foxmail.com

features using the 2D encoder (see Section "Biomedical CLIP Vision Encoder"), followed by filtering redundant slices via a cosine similarity strategy¹¹ (see Section "Cosine Similarity based Slice Selection"). Subsequently, a spatial-frequency cross-domain fusion module reconstructs the 3D context and enriches feature dimensions (see Section "Spatial-Frequency Cross-domain Feature Fusion"). Finally, a fine-grained feature injection mechanism preserves critical diagnostic details while efficiently compressing features (see Section "Fine-grained Feature Injection Mechanism"), providing an efficient and information-dense visual input for the Large Language Model's final reasoning. Experiments show that on the M3D-Cap report generation task, our model achieves a METEOR score¹² of 50.13, an improvement of 37.6% over the previous state-of-the-art model (Med3DVLM⁶); in the M3D-VQA closed-ended task, it achieves an average accuracy² of 82.90%, leading by 3.15 percentage points. Our proposed data-efficient framework is capable of achieving state-of-the-art performance on key benchmarks without the need for a pre-trained 3D vision encoder.

In summary, our contributions are as follows:

- We propose a data-efficient framework for 3D medical image analysis that uniquely couples a pre-trained 2D encoder with specialized structural enhancement modules. This method constructs a 3D visual representation by reorganizing and enhancing features from 2D slices, thereby enabling the use of a pre-trained 2D encoder for 3D vision-language tasks and circumventing the dependency on large-scale 3D datasets.
- We design a collaborative multi-module feature optimization mechanism that strategically retains representative visual features via slice selection, and fundamentally reconstructs 3D semantic coherence through a novel Spatial-Frequency Fusion. This architecture, further enhanced by a fine-grained injection mechanism for high-fidelity detail preservation, effectively resolves the intrinsic limitations of 2D encoders without the need for 3D pre-training.
- We conducted a comprehensive evaluation, the results of which show that our model's performance comprehensively surpasses that of previous methods on multiple benchmarks, including M3D-Cap for report generation and M3D-VQA for visual question answering, validating the superiority of the proposed framework.

Related work

Medical multimodal large language models

Medical Visual-Language Models (Med-VLM)¹ aim to achieve deep semantic alignment between medical images and textual information, supporting tasks such as Medical Visual Question Answering (VQA)¹³, lesion localization¹⁴, and automatic generation of medical reports¹⁵. This helps improve the accuracy and consistency of diagnostic reports and effectively alleviates the workload of clinicians. Early research¹⁶ primarily employed joint pre-training of visual and language modalities, using Encoder-only architectures, which initially endowed the model with the ability to understand cross-modal information and perform basic text generation. With the development of Large Language Models (LLMs)^{17–20}, Med-VLMs have integrated the powerful reasoning and generalization capabilities of LLMs, leading to a series of representative models^{21–23} that have achieved significant performance improvements in 2D medical image-based question answering and report generation tasks. Furthermore, to better meet the clinical demand for processing 3D medical images, recent models^{4–6} have focused on modeling 3D medical visual information and integrating multimodal data. This allows the model to understand 3D spatial structures and recognize lesion details, further enhancing the model's generalization ability, diagnostic accuracy, and clinical interpretability.

Despite these advancements, current 3D medical visual-language models are generally highly dependent on specially designed 3D vision encoders (such as 3D Vision Transformers or 3D-CLIP)^{4–6}. They require large-scale, high-quality 3D medical image data. However, since 3D medical data are far less available than 2D data, training vision encoders directly on 3D data faces the problem of data scarcity. This not only increases the cost of data annotation and computational overhead but also limits the model's generalization ability and practical application potential.

3D medical image feature extraction and modeling

3D medical images are presented as volumetric data, providing rich spatial structures and lesion detail information. However, 3D image data has high dimensionality and dense slices, leading to problems of visual feature redundancy, which poses challenges for existing feature extraction methods. In recent years, researchers have proposed a series of 3D visual feature modeling approaches. One approach, such as^{4–6}, directly extracts 3D visual features using 3D convolutions or 3D Transformers. However, these methods usually require large amounts of high-quality 3D data for training, which involves high data annotation and computational costs. Another approach, such as^{24,25}, utilizes pre-trained 2D vision encoders to extract slice features layer by layer. Although this strategy alleviates the dependency on 3D data, it introduces new problems. First, this slice-by-slice independent processing method naturally ignores the spatial correlations and structural continuity between slices, which is crucial for understanding the complete 3D context. Second, the problem of visual feature redundancy between slices is still prominent. Existing work often employs simple pooling operations to reduce the number of visual tokens, but this rough downsampling method can easily lose key lesion details and lacks effective strategies for cross-slice feature fusion. Furthermore, most mainstream methods focus primarily on modeling spatial domain features. While this can effectively capture the intuitive anatomical structure of lesions, it performs poorly in sensing the intrinsic texture differences and structural heterogeneity within lesions.

Therefore, how to improve data efficiency while effectively reducing visual feature redundancy, and fully utilize the spatial correlations and visual representation potential between slices, remains a key problem that needs to be addressed in the field of 3D medical image analysis.

Method

As shown in Figure 1, our method consists of four main components: a biomedical 2D vision encoder, a spatial-frequency cross-domain feature fusion module, a fine-grained feature injection mechanism, and a large language model (LLM), for performing 3D medical vision-language tasks. First, given input 3D medical image data, we utilize a pretrained biomedical-specific 2D vision encoder to extract features from different slices. Then, redundant slices are filtered by cosine similarity matching, retaining representative slice features. Subsequently, the spatial-frequency cross-domain feature fusion module is employed to fuse spatial and frequency domain features, generating a comprehensive representation containing both intra-slice details and inter-slice global correlations. Afterwards, these features are downsampled to obtain a low-resolution representation. To further enhance the representation ability of the low-resolution features, we design a fine-grained feature injection mechanism, which injects high-resolution features into low-resolution features and updates them within local contexts, thus enhancing the detailed information of the low-resolution representation. Finally, the enhanced low-resolution visual tokens, together with language embeddings, are fed into the large language model for task-specific semantic reasoning, producing task-related answers.

Biomedical CLIP vision encoder

When processing 3D medical images, directly utilizing 3D vision encoders to extract visual features faces significant challenges, mainly including data scarcity and high computational costs. Data scarcity limits the feature extraction and generalization capabilities of 3D encoders, whereas 2D encoders exhibit better generalization performance due to richer and more easily accessible training data. In addition, training 3D vision encoders requires substantial computational resources, leading to expensive computational costs, which restrict their applicability in medical imaging.

Therefore, we adopt a pretrained 2D biomedical vision encoder, specifically BioMedCLIP⁹. BioMedCLIP is specially designed for medical images and pretrained on large-scale medical datasets, enabling efficient extraction of fine-grained visual features. By splitting each medical image into multiple panel images and subsequently constructing the PMC-Fine-Grained-46M dataset, BioMedCLIP can better adapt to the high resolution and complex structures of medical images, extracting more detailed visual features. When processing 3D medical image slices, BioMedCLIP effectively captures details such as tissue structures, textures, and anatomical morphology.

Specifically, given the spatial resolution of input medical imaging data as $H \times W$, we slice the 3D image along the depth direction (e.g., axial, coronal, or sagittal) into a series of continuous 2D slices, with the number of slices (depth dimension) denoted as D . Each slice, treated as an independent 2D image, is input to the pretrained BiomedCLIP_256-vit_base_patch16_224 vision encoder for feature extraction. After being processed by the BioMedCLIP encoder, each slice is represented as a feature matrix:

$$F_i \in \mathbb{R}^{N \times C}, \quad i = 1, 2, \dots, D \quad (1)$$

where N denotes the number of patches per slice (e.g., $N = 14 \times 14 = 196$), and C is the feature dimension of each patch (e.g., $C = 768$). After processing all D slices, the preliminary visual feature representation for the entire 3D medical imaging data is:

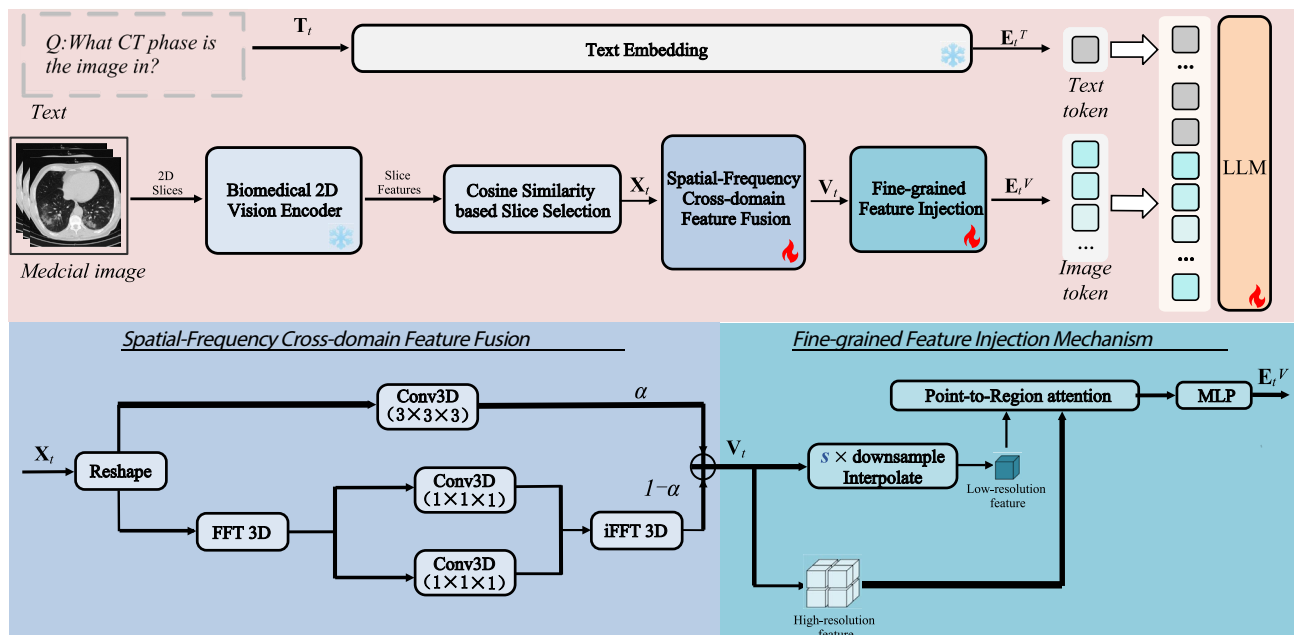


Fig. 1. Overview of the proposed data-efficient 3D medical vision-language model framework.

$$F \in \mathbb{R}^{D \times N \times C} \quad (2)$$

This approach circumvents the high data volume and computational cost required for training 3D vision encoders, while extracting finer intra-slice features.

Cosine similarity based slice selection

Due to the inherent visual redundancy between adjacent slices in 3D medical images, directly processing all slices would lead to unnecessary computational resource consumption. To precisely and efficiently remove redundant slices, we propose a cosine similarity²⁶ based slice selection strategy.

Specifically, given the visual features extracted from the first stage as $F \in \mathbb{R}^{D \times N \times C}$, where D , N , and C denote the number of input slices, the number of patches per slice, and the feature dimension of each patch respectively, we first obtain the global feature vector f_i for each slice through global average pooling:

$$f_i = \frac{1}{N} \sum_{j=1}^N F_i(j), \quad f_i \in \mathbb{R}^C \quad (3)$$

Then, we compute the cosine similarity $S(f_i, f_{i+1})$ between adjacent slices:

$$S(f_i, f_{i+1}) = \frac{f_i \cdot f_{i+1}}{\|f_i\| \|f_{i+1}\|} \quad (4)$$

Based on the computed similarities, we sort all slices and directly retain the top $\rho\%$ slices with the lowest similarity scores. This explicitly controls the number of slices for subsequent processing, finally yielding a reduced feature representation:

$$\hat{F} \in \mathbb{R}^{\hat{D} \times N \times C}, \quad \hat{D} = \lfloor D \times \rho\% \rfloor \quad (5)$$

This strategy proactively reduces redundancy in slice features in a flexible and controllable manner, significantly improving the efficiency of subsequent model processing.

Spatial-frequency cross-domain feature fusion

After slice selection based on cosine similarity, we obtain a reduced visual feature representation:

$$\hat{F} \in \mathbb{R}^{\hat{D} \times N \times C} \quad (6)$$

We first reshape the features into an explicit spatial-depth 3D representation to capture structural information along spatial and depth dimensions:

$$\hat{F}_{reshape} \in \mathbb{R}^{\hat{D} \times H' \times W' \times C} \quad (7)$$

where \hat{D} , H' , W' , C denote the reduced number of slices, spatial height, spatial width, and feature dimension, respectively.

To restore the spatial-depth correlations among slices and further enhance feature representation, drawing inspiration from recent feature enhancement strategies in lightweight medical models²⁷, we design a spatial-frequency cross-domain fusion mechanism to fully leverage the complementary information from both domains.

Spatial-domain Path

We apply 3D convolution²⁸ (kernel size $3 \times 3 \times 3$) directly on reshaped features to capture local spatial structural information:

$$F_{spa} = \text{Conv3D}_{3 \times 3 \times 3}(\hat{F}_{reshape}), \quad F_{spa} \in \mathbb{R}^{\hat{D} \times H' \times W' \times C} \quad (8)$$

3D convolution has an explicit local receptive field along spatial and depth dimensions, enabling sensitive capturing of detailed structures, edges, and texture features.

Frequency-domain Path

To further capture global and long-range structural information beyond the reach of spatial convolutions, we adopt frequency-domain feature extraction to explicitly capture global slice-wise features. Specifically, we first transform spatial-depth features into frequency domain via 3D Fast Fourier Transform²⁹ (3D FFT):

$$F_{freq}^{real}, F_{freq}^{imag} = \text{FFT3D}(\hat{F}_{reshape}) \quad (9)$$

Then, we refine the real and imaginary parts of frequency-domain features via 3D convolution (kernel size $1 \times 1 \times 1$):

$$\tilde{F}_{freq}^{real} = \text{Conv3D}_{1 \times 1 \times 1}(F_{freq}^{real}), \quad \tilde{F}_{freq}^{imag} = \text{Conv3D}_{1 \times 1 \times 1}(F_{freq}^{imag}) \quad (10)$$

Next, we restore refined frequency-domain features back to the spatial domain via inverse FFT³⁰ (iFFT3D):

$$F_{freq} = \text{iFFT3D}(\tilde{F}_{freq}^{real}, \tilde{F}_{freq}^{imag}), \quad F_{freq} \in \mathbb{R}^{\hat{D} \times H' \times W' \times C} \quad (11)$$

Frequency-domain convolution naturally has a non-local receptive field, effectively capturing global texture and inter-slice structural correlations, thus compensating for the loss of cross-slice information caused by slice-wise independent feature extraction.

Adaptive Spatial-Frequency Fusion

We dynamically generate fusion weights α :

$$\alpha = \sigma(\text{Conv3D}_{1 \times 1 \times 1}(F_{spa} \oplus F_{freq})) \quad (12)$$

where \oplus denotes feature concatenation, and $\sigma(\cdot)$ is the sigmoid activation function³¹.

The final cross-domain fused features are given by:

$$F^{SF} = \alpha \cdot F_{spa} + (1 - \alpha) \cdot F_{freq}, \quad F^{SF} \in \mathbb{R}^{\hat{D} \times H' \times W' \times C} \quad (13)$$

This adaptive fusion enhances the comprehensiveness and robustness of final feature representations. Critically, traditional slice-stacking methods treat 2D slices as independent instances, inevitably severing the volumetric continuity along the Z-axis. This leads to a “semantic gap” where the model fails to perceive the 3D organ geometry. Our module acts as a necessary remedy by explicitly modeling inter-slice dependencies through the frequency domain. Since frequency components naturally capture global repetitive patterns across the entire volume, this design effectively restores the 3D spatial consistency lost during 2D encoding, ensuring the model perceives the medical volume as a continuous whole rather than a disjointed collection of images.

Fine-grained feature injection mechanism

After the spatial-frequency cross-domain fusion mechanism, we obtain the enhanced spatial-frequency feature representation as:

$$F^{SF} \in \mathbb{R}^{\hat{D} \times H' \times W' \times C} \quad (14)$$

To further reduce the computational overhead in subsequent processing while retaining the critical structural and detail information in medical images, we design two output paths, generating high- and low-resolution feature representations:

- **High-resolution path:** Directly retains the original spatial dimensions of the fused features to preserve feature details.
- **Low-resolution path:** Downsamples the high-resolution features by a factor of s using trilinear interpolation³² to reduce the computational burden.

The specific representations are as follows:

$$F_{high} = F^{SF} \in \mathbb{R}^{\hat{D} \times H' \times W' \times C}, \quad F_{low} = \text{TrilinearInterp}(F^{SF}) \in \mathbb{R}^{\frac{\hat{D}}{s} \times \frac{H'}{s} \times \frac{W'}{s} \times C} \quad (15)$$

Although the low-resolution path significantly reduces the computational complexity, some structural detail information is inevitably lost. **In general vision tasks, such loss may be acceptable; however, in medical diagnostics, it is critical, as key diagnostic indicators often occupy extremely small regions within the volumetric data. Traditional pooling-based downsampling strategies, by their nature, reduce spatial resolution and inevitably discard high-frequency fine-grained details, rendering these small anomalies undetectable. Therefore, our cross-scale feature injection mechanism is a necessity. Unlike simple downsampling, it establishes a detail-preserving bridge that selectively re-integrates high-resolution features into the visual tokens, allowing the model to recover lost details in the low-resolution path while maintaining computational efficiency.**

Specifically, we propose a Query-Key-Value³³ based Point-to-Region attention mechanism to implement the cross-scale feature injection process. Each spatial location feature in the low-resolution path is treated as an independent point query (Point Query) to retrieve detailed information from the corresponding region in the high-resolution path. Specifically, we project the low-resolution features F_{low} as Queries, and project the high-resolution features F_{high} as Keys and Values:

$$Q = W_Q(F_{low}), \quad K = W_K(F_{high}), \quad V = W_V(F_{high}) \quad (16)$$

Here, W_Q, W_K, W_V are independent linear mappings to ensure that the feature dimensions of Q, K , and V are suitable for the attention mechanism calculation. To facilitate the attention calculation, we reshape the projected features into a two-dimensional matrix form with dimensions as follows:

$$Q \in \mathbb{R}^{\left(\frac{\hat{D}}{s} \times \frac{H'}{s} \times \frac{W'}{s}\right) \times C}, \quad K, V \in \mathbb{R}^{(\hat{D} \times H' \times W') \times C} \quad (17)$$

Then, we calculate the cross-scale attention weights to determine the relationship between the low-resolution features and high-resolution features. First, the Q, K , and V features are reshaped into two-dimensional matrices, and the attention weights are computed:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{C}}\right)V \quad (18)$$

The attention mechanism naturally captures long-range, cross-scale relationships between the low-resolution and high-resolution features, thus effectively injecting the detail information from the high-resolution features into the low-resolution features.

After the attention feature injection, the features are restored to the original dimensions of the low-resolution path and further updated through a simple multi-layer perceptron³⁴ (MLP) for feature mapping and fusion:

$$F_{\text{enhanced}} = \text{MLP}(\text{Attention}(Q, K, V)), \quad F_{\text{enhanced}} \in \mathbb{R}^{\frac{\hat{D}}{s} \times \frac{H'}{s} \times \frac{W'}{s} \times C} \quad (19)$$

Finally, we obtain the enhanced low-resolution feature representation F_{enhanced} , which effectively restores the key details lost in the low-resolution path without significantly increasing computational load, further improving the comprehensiveness and robustness of the feature representation.

Tokenization for LLM model

To input the low-resolution features into the large language model (LLM), we first flatten the low-resolution features $F_{\text{enhanced}} \in \mathbb{R}^{\frac{\hat{D}}{s} \times \frac{H'}{s} \times \frac{W'}{s} \times C}$ obtained from the fine-grained feature injection mechanism. The flattening operation expands each spatial dimension $\frac{\hat{D}}{s} \times \frac{H'}{s} \times \frac{W'}{s}$, while keeping the feature channel C unchanged. We use M' to represent the flattened spatial dimensions:

$$M' = \frac{\hat{D}}{s} \times \frac{H'}{s} \times \frac{W'}{s} \quad (20)$$

The flattened visual feature representation is:

$$F_{\text{flattened}} \in \mathbb{R}^{M' \times C} \quad (21)$$

This flattening operation ensures that the spatial information of the low-resolution features $\frac{\hat{D}}{s} \times \frac{H'}{s} \times \frac{W'}{s}$ is effectively laid out and is compatible with the input format of the language model.

When processing textual input, we use the pretrained tokenizer of the large language model (LLM) to tokenize the text sequence and project it into the text embedding space. The visual and text embeddings will be arranged in sequence and input into the LLM for understanding. To enhance the semantic representation of the image, we add a prefix prompt: "< image> n This provides an overview of the picture." where <image> is a special token that will later be replaced by the image-level embedding extracted by the fine-grained feature injection mechanism.

Compared to the initial 3D medical image slices, we achieve a significant reduction in the number of visual tokens, with a reduction ratio of $\rho\% \times \left(\frac{1}{s^3}\right)$, reflecting our efficiency improvement.

Experiments

Our method directly utilizes a pre-trained 2D vision encoder, bypassing the pre-training step of a 3D vision encoder, significantly improving the data efficiency of 3D medical image analysis. In the experiments, we first demonstrate our advantage in data efficiency by comparing with the latest model architectures and the required amount of training data. Building on this advantage, we focus on evaluating the performance improvements of our method in medical visual question answering and report generation tasks compared to existing state-of-the-art methods. Figure 2 shows the visualization results. Finally, through ablation studies, we explore the optimal path of our method and the contribution of each module.

Task and datasets

This study focuses on two key tasks in 3D medical vision-language understanding: *Medical Report Generation* and *Medical VQA*. We use the publicly available large-scale 3D multimodal dataset, M3D-Data, for model training and evaluation. This dataset covers various anatomical structures and lesion types, providing rich medical imaging data along with corresponding textual annotations, ensuring good representativeness and wide applicability. To ensure fairness in our experiments and to facilitate comparison with existing studies, we adopt two subsets of the M3D-Data dataset⁴, as used in previous research:

- **M3D-Cap:** This subset consists of image-text pairs for medical report generation. The training set includes 115 K image-caption pairs, while the test set contains 2 K pairs. The primary task is to generate accurate and coherent medical reports based on the given medical images.
- **M3D-VQA:** This subset contains image-question pairs for the medical visual question answering task. The training set consists of approximately 420 K open-ended and closed-ended question-answer pairs, while the test set includes around 13 K pairs. The task requires the model to infer the correct answer based on 3D medical images and corresponding question texts, covering core medical topics such as lesion location, organ recognition, and abnormality detection.

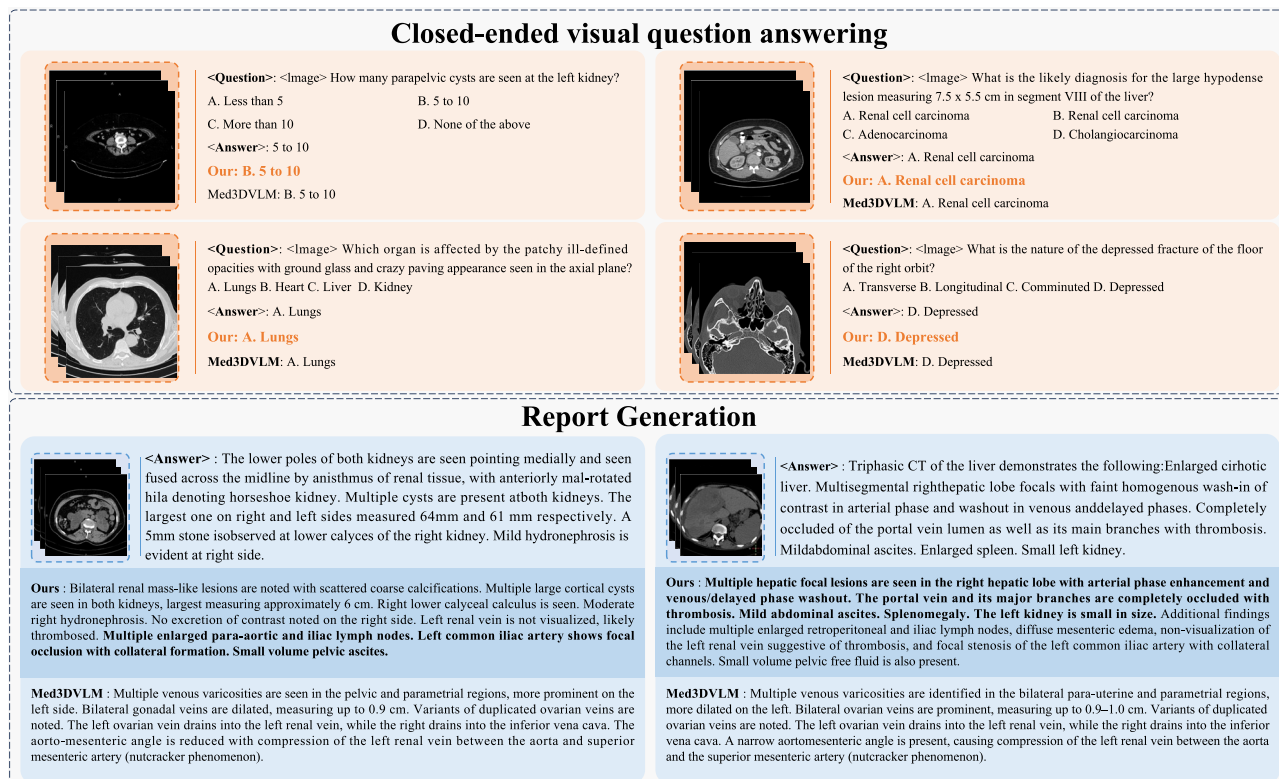


Fig. 2. Examples of VQA and report generation results from Med3DVLM and our model.

Implementation details

We apply Min-Max Normalization³⁵ consistently to preprocess the 3D CT images as input. Additionally, we resize and crop the 3D images to a standardized size of $50 \times 224 \times 224$. We use a pretrained biomedical-specific 2D vision encoder, BioMedCLIP, with the configuration of `vit_base_patch16_224` to extract features from the 3D medical image slices. During slice selection, we set the ratio coefficient $\rho\% = 28\%$. This specific threshold was empirically determined via a grid search³⁶ on the validation set, where 28% yielded the optimal trade-off between computational efficiency and feature retention. In the fine-grained feature injection mechanism, we use a downsampling factor $s = 2$. Furthermore, to ensure fair comparison with most models, we also use LLaMA-2-7B³⁷ as the language model.

The training process is divided into two stages:

- **First stage:** We freeze the vision encoder and LLM, and fine-tune the remaining modules, using image-text pairs for training. The batch size is set to 12×8 , the learning rate³⁸ is 10^{-4} , with warm-up³⁹ and cosine decay strategies⁴⁰ applied.
- **Second stage:** We freeze the vision encoder and jointly fine-tune the remaining modules using medical image descriptions and question-answer data. The batch size is set to 12×8 , the learning rate is adjusted to 2×10^{-5} , and we continue to use warm-up and cosine decay strategies. In this stage, we employ the LoRA⁴¹ (Low-Rank Adaptation) method to fine-tune the language model's parameters. The LoRA parameters are set to $r = 16$, $\alpha = 32$, with a dropout⁴² rate of 0.1, and the maximum context length is set to 512 to further improve fine-tuning efficiency.

All training is performed using the AdamW optimizer⁴³, with acceleration provided by the DeepSpeed framework⁴⁴ and the use of bf16 mixed-precision⁴⁵ training. The training process is parallelized across four NVIDIA A800 GPUs, each with 80GB of memory.

Evaluation metrics

For evaluating the quality of medical report generation, we use the following automated metrics:

- **BLEU**⁴⁶: This metric evaluates the accuracy of the generated report by calculating the n-gram overlap between the generated text and the reference text.
- **ROUGE**⁴⁷: This metric measures the recall between the generated text and the reference text, reflecting how much of the reference content is covered by the generated text.
- **METEOR**¹²: An improvement over BLEU and ROUGE, METEOR considers synonym matching and stemming, providing a balance between precision and recall.

Model Name	Architecture
RadFM ⁵	3D Vision Transformer (ViT) Encoder + Perceiver Module + MedLLaMA-13B LLM
M3D-LaMed ⁴	3D Vision Transformer (ViT) Encoder + 3D Spatial Pooling Perceiver + LLaMA-2-7B LLM
Med3DVLM ⁶	DCFormer Vision Encoder + Dual-Stream MLP-Mixer Projector + Qwen2.5-7B-Instruct LLM
Ours	2D Vision Encoder + Feature Enhancement and Token Compression + LLaMA-2-7B LLM

Table 1. Architectural comparison of our proposed model with baseline 3D medical vision-language models.

Model	BLEU	ROUGE	METEOR	BERT-Score
RadFM ⁵	12.23	16.49	11.57	87.93
M3D-LaMed ⁴	15.15	19.55	14.38	88.46
Med3DVLM ⁶	36.88	40.25	36.42	88.11
Ours	50.26	53.37	50.13	89.98

Table 2. Comparison of report generation performance on the M3D-Cap dataset.

- **BERTScore**⁴⁸: Using contextual embeddings from BERT, this metric evaluates the semantic similarity between the generated text and the reference text.

These metrics comprehensively evaluate the quality of medical report generation in terms of precision, recall, and semantic consistency.

For VQA, we adopt the following evaluation criteria:

- **Accuracy**²: Used for closed-ended VQA, this metric measures the proportion of correctly chosen answers from predefined options.
- **BLEU**⁴⁶, **ROUGE**⁴⁷, **METEOR**¹², **BERTScore**⁴⁸: These metrics are used for open-ended VQA to assess the similarity between the generated answer and the reference answer.

These metrics comprehensively assess the model's performance in VQA tasks, considering both the correctness of the answers and the quality and semantic consistency of the generated responses.

Baseline comparison

As shown in Table 1, the state-of-the-art model architectures are listed. Compared to existing models, our approach avoids the pretraining of 3D vision encoders. The 2D vision encoder (e.g., BioMedCLIP), pretrained on large-scale internet data, captures medical image details more effectively, especially given the scarcity of 3D data. Our model aims to achieve a good balance between cost and performance, fully leveraging the advantages of 2D vision encoders in medical image analysis. Moreover, while other methods rely on coarse pooling⁴⁹ to reduce the number of visual tokens, we introduce a fine-grained feature injection mechanism that injects more critical semantic information during compression.

Evaluation on report generation

In the report generation task, our model demonstrates superior performance across all evaluation metrics. As shown in Table 2, our model achieves scores of 50.26, 53.37, and 50.13 for BLEU, ROUGE, and METEOR, respectively, showing a significant improvement compared to RadFM and M3D-LaMed, particularly with a 35% increase in METEOR. This result indicates that our model generates semantically rich and content-relevant medical reports with substantial improvements.

Although the improvement in BERT-Score is relatively small, our model still achieves a score of 89.98, which is higher than that of the other models. This suggests that our model excels in semantic alignment, with the generated reports being more precise and fluent in both language expression and semantic consistency.

These improvements can be attributed to the design of our method, particularly the use of a 2D vision encoder trained on a significantly larger dataset compared to the 3D vision encoders used by other models. This allows our model to more effectively capture the details in medical images, thereby enhancing the quality and accuracy of report generation. Additionally, the fine-grained feature injection mechanism injects high-resolution details into low-resolution features, further improving the semantic details and structural coherence during report generation.

Qualitative Analysis. To further validate the model's capability in capturing nuanced medical details, we conducted a qualitative comparison as shown in Figure 2. In the first case (kidney analysis), our model successfully identified subtle pathological features such as “scattered coarse calcifications” and “mild hydronephrosis,” whereas the baseline Med3DVLM failed to capture these fine-grained texture details, resulting in a generic description. Similarly, in the second case (liver tumor diagnosis), our model accurately described the specific hemodynamic characteristics, including “arterial phase enhancement” and “venous/delayed phase washout.” This demonstrates that our Spatial-Frequency Fusion and Fine-grained Injection mechanisms effectively preserve critical diagnostic information that is typically lost in traditional downsampling processes, thereby generating reports that are clinically more accurate and valuable.

Open-ended VQA							
Method	Metric	Plane	Phase	Organ	Abnormality	Location	Mean
RadFM ⁵	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 ⁴	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 ⁶	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
Ours	BLEU	99.07	82.71	49.82	26.79	32.61	58.20
	ROUGE	99.15	89.32	52.39	29.37	38.79	61.80
	METEOR	52.32	75.91	40.06	25.81	28.19	44.46
	BERTScore	99.87	97.13	93.75	91.27	91.64	94.73

Table 3. Comparison of open-ended VQA performance on the M3D-VQA dataset.

Close-ended VQA						
Methods	Plane	Phase	Organ	Abnormality	Location	Mean
RadFM ⁵	19.65	28.70	16.80	18.92	14.88	19.79
M3D-LaMed ⁴	98.80	79.75	74.75	66.65	58.94	75.78
Med3DVLM ⁶	99.15	87.50	77.45	70.17	64.49	79.75
Ours	99.82	90.04	79.93	74.41	70.32	82.90

Table 4. Comparison of closed-ended VQA accuracy on the M3D-VQA dataset.

Evaluation on VQA

In the VQA task, our model demonstrates exceptional performance on both open-ended and closed-ended questions, outperforming existing baseline models.

As shown in Table 3, in the open-ended VQA task, our model achieved average scores of 58.20 for BLEU, 61.80 for ROUGE, 44.46 for METEOR, and 94.73 for BERTScore, significantly surpassing advanced models such as Med3DVLM. Particularly, in complex categories that require deep semantic understanding and fine-grained detail perception, such as “Organ,” “Abnormality,” and “Location,” our model’s advantages are especially pronounced. For instance, in the “Abnormality” category, our ROUGE score (29.37) improved by over 26% compared to Med3DVLM (23.27), indicating that the answers generated by our model are closer to the ground truth and better at identifying and describing lesion features.

Similarly, in the closed-ended VQA task shown in Table 4, our model achieved a new state-of-the-art (SOTA) with an average accuracy of 82.90%, outperforming the next best model, Med3DVLM, by 3.15 percentage points. The model consistently led across all subtask categories, especially in the most challenging categories, “Abnormality” and “Location,” where accuracy reached 74.41% and 70.32%, respectively, demonstrating its strong reasoning and localization capabilities.

The superior performance of our model can be attributed to its unique visual processing pipeline. This approach effectively captures fine image details and, through efficient feature fusion and compression strategies, provides a compact visual representation containing key diagnostic information to the large language model, thereby improving performance while ensuring computational efficiency.

Evaluation on efficiency

We first evaluate the visual encoder strategies and associated data costs. As shown in Table 5, mainstream 3D models (e.g., RadFM⁵ and M3D-LaMed⁴) necessitate massive datasets (ranging from 120k to 16M samples) to re-train their 3D visual encoders. In contrast, by leveraging a frozen 2D BioMedCLIP, our method completely eliminates the need for 3D visual encoder re-training.

Furthermore, we evaluate the computational resource consumption during the end-to-end fine-tuning stage (excluding the pre-training and alignment of the visual encoder). Since RadFM utilizes web-scale data for pre-training, we limit its evaluation to inference latency in this context. As presented in Table 6, our model achieves optimal efficiency in both training time and inference speed. This superior performance is primarily attributed to our utilization of a 2D vision encoder, which has significantly lower computational complexity compared to 3D encoders. Additionally, RadFM exhibits notably higher inference latency (267 ms), which results from

Model	Vision Encoder	3D Encoder Re-training	3D Pre-training Data
RadFM ⁵	3D ViT	Yes	~16 M
M3D-LaMed ⁴	3D ViT	Yes	120 k
Med3DVLM ⁶	DCFormer (3D)	Yes	120 k
Ours	2D BioMedCLIP	No	0

Table 5. Comparison of vision encoder strategies and data dependencies.

Model	Training Time (GPU Hours)	Inference Latency (ms)
RadFM ⁵	-	267
M3D-LaMed ⁴	232	165
Med3DVLM ⁶	218	154
Ours	212	146

Table 6. Comparison of training efficiency and inference latency.

Model	MRI (Acc %)	PET (Acc %)
RadFM ⁵	85.4	82.1
M3D-LaMed ⁴	72.6	69.8
Med3DVLM ⁶	74.1	71.5
Ours	93.8	91.5

Table 7. Comparison of accuracy on MRI and PET modalities.

its use of a larger 13B language model, whereas other comparative models (including ours) are built upon 7B architectures.

Evaluation of cross-modal adaptability

To validate the generalizability of our framework to other 3D imaging modalities (such as MRI and PET), we constructed a small-scale multi-modal diagnostic benchmark dataset. We first randomly selected 1,000 3D medical image samples drawn from BraTS 2023⁵⁰ (MRI, $N = 500$) and AutoPET⁵¹ (PET, $N = 500$). Subsequently, a Chief Radiologist with over 10 years of experience constructed closed-ended VQA pairs for each case, covering three dimensions: “Modality Identification”, “Lesion Detection”, and “Anatomical Localization”, resulting in a total of 3,000 VQA pairs.

To ensure the rigor of the evaluation benchmark, we invited three additional independent Chief Radiologists to serve as evaluators. Each evaluator independently reviewed all samples to verify the correctness of the VQA pairs and their consistency with the imaging data. The initial round of evaluation yielded a unanimous acceptance rate of 99.67%. Any sample receiving a dissenting vote (i.e., failing to reach 100% consensus) was immediately discarded and replaced with a new sample extracted from the source dataset for re-evaluation, until all samples in the dataset achieved 100% expert approval. To quantify the reliability of the evaluation process, we employed the Inter-Annotator Agreement (IAA)⁵² statistical method. The Cohen’s Kappa coefficient, calculated based on the independent review results, was 0.927, indicating “near-perfect agreement” among the experts. Finally, the quality-assured dataset was randomly partitioned into a training set and a testing set with a ratio of 8:2. All 3D volumetric data were preprocessed following the protocols detailed in Section 4.2.

We report the model accuracy on the test set in Table 7. The results demonstrate that our model achieves superior adaptability, significantly outperforming baselines with accuracies of 93.8% (MRI) and 91.5% (PET). This performance advantage primarily stems from the inherent robustness of our pre-trained 2D vision encoder. Unlike specialized 3D encoders that are often constrained by the scarcity of modality-specific volumetric data, our 2D encoder inherits generalized visual representations from massive-scale 2D biomedical pre-training. This allows our framework to effectively bridge the domain gap and rapidly generalize to diverse imaging modalities with minimal fine-tuning, validating the data-efficiency of our approach.

Ablation study

We conducted a series of ablation studies to assess the contribution of each key component in our proposed model, as demonstrated in Table 8.

Impact of the Vision Encoder. Replacing our BioMedCLIP with another in-domain encoder, PubMedCLIP¹⁰, leads to a slight degradation in performance (81.95% VQA accuracy), whereas using a general-domain encoder, CLIP-ViT-L/14⁵³, results in a substantial drop (75.90% VQA accuracy). This confirms that domain-specific pre-training is critical for extracting precise medical visual representations.

Method	Report Generation				Medical VQA				
	BLEU	ROUGE	METOR	BERT	BLEU	ROUGE	METOR	BERT	Accuracy
Ours	50.26	53.37	50.13	89.98	58.20	61.80	44.46	94.73	82.90
PubMedCLIP	48.92	51.88	48.55	89.51	56.95	60.12	42.87	94.15	81.65
CLIP-ViT-L/14	41.15	44.23	40.89	88.04	51.33	54.08	38.16	92.88	77.21
Uniform Sampling	48.17	51.02	48.01	89.46	56.51	59.83	42.30	93.99	81.12
Spatial Path Only	47.95	50.81	47.63	89.29	56.02	59.44	41.88	94.03	82.01
Frequency Path Only	45.33	48.16	44.92	88.75	54.18	57.25	39.95	93.17	80.76
No Injection Module	46.58	49.20	46.11	89.01	54.88	57.99	40.52	93.64	80.48

Table 8. Ablation study evaluating the contribution of each key component in our framework.

Impact of the Cosine Similarity-based Slice Selection Strategy. Our cosine similarity-based selection strategy outperforms uniform sampling, with respective VQA accuracies of 82.90% and 80.57%. This demonstrates that our proposed method effectively filters out redundancy while better preserving critical diagnostic features.

Impact of the Spatial-Frequency Cross-domain Feature Fusion. The fusion module is designed to establish 3D spatial correlations from 2D slice features. Relying on a single path degrades performance. Specifically, the frequency-domain path preserves global distribution and long-range dependencies at the cost of high-frequency spatial information, which reduces local feature discriminability. Conversely, the spatial-domain path excels at preserving high-frequency local details via its local receptive field but is limited in modeling global context, resulting in sub-optimal performance. The results confirm that an optimal 3D representation requires a synthesis of the local structural fidelity from the spatial domain and the global contextual awareness from the frequency domain.

Impact of the Fine-grained Feature Injection Mechanism. Removing the fine-grained injection module leads to a significant performance drop (78.92% VQA accuracy), highlighting the criticality of information fidelity during feature compression. Under the No Injection Module configuration, direct downsampling via trilinear interpolation acts as a low-pass filter, causing feature homogenization where highly discriminative, high-frequency details are smoothed out, thereby reducing the overall information density. Our attention-based injection mechanism counteracts this by allowing the compressed, low-resolution tokens to selectively aggregate salient information from the high-resolution feature map. This process ensures that the final token sequence, while compact, retains the high information fidelity essential for complex reasoning tasks.

Conclusion

In this paper, we proposed a data-efficient framework for 3D medical vision-language tasks that circumvents the need for specialized 3D encoders by leveraging a pre-trained 2D vision encoder. Our pipeline strategically refines 2D slice features through redundancy pruning, spatial-frequency fusion to restore 3D context, and a fine-grained injection mechanism to preserve diagnostic details. Extensive evaluations on public benchmarks demonstrate that our model achieves state-of-the-art performance in both report generation and visual question answering. Our work presents a scalable and effective method for 3D medical multimodal analysis, offering a viable alternative to data-intensive 3D encoders and contributing to the development of more efficient and adaptable multimodal clinical systems.

Limitations

Despite the promising results demonstrated by our framework, this study has several limitations that merit discussion. First, our current model is trained and evaluated exclusively on English-language medical data. Given that clinical workflows often involve multi-lingual reporting, extending the model's capability to support other languages remains a critical direction for future work. Second, while we have validated the model's adaptability to MRI and PET modalities, we have not yet evaluated its performance on other diverse imaging types, such as Ultrasound or Optical Coherence Tomography (OCT). Finally, although our data-efficient approach mitigates the dependency on massive datasets, the overall scale of publicly available 3D medical-caption pairs is still significantly smaller than that of 2D generic vision-language datasets. We plan to address these limitations in future research by expanding the diversity of training data and integrating multi-lingual instruction tuning.

Data availability

The datasets utilized in this study are publicly available. The primary training and evaluation datasets, M3D-Data (including its subsets M3D-Cap and M3D-VQA), are accessible via their official repository. Additionally, the datasets used for the cross-modal adaptability evaluation, BraTS 2023 and AutoPET, are publicly available through their respective challenge platforms.

Received: 4 September 2025; Accepted: 5 February 2026

Published online: 13 February 2026

References

- Hartsock, I. & Rasool, G. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence* 7, 1430984 (2024).

2. Antol, S. et al. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433 (2015).
3. Wang, J., Bhalerao, A. & He, Y. Cross-modal prototype driven network for radiology report generation. In *European Conference on Computer Vision*, 563–579 (Springer, 2022).
4. Bai, F., Du, Y., Huang, T., Meng, M. Q.-H. & Zhao, B. M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint [arXiv:2404.00578](https://arxiv.org/abs/2404.00578) (2024).
5. Wu, C. et al. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications* **16**, 7866 (2025).
6. Xin, Y., Ates, G. C., Gong, K. & Shao, W. Med3dvlm: An efficient vision-language model for 3d medical image analysis. arXiv preprint [arXiv:2503.20047](https://arxiv.org/abs/2503.20047) (2025).
7. Lahoud, J. et al. 3d vision with transformers: A survey. arXiv preprint [arXiv:2208.04309](https://arxiv.org/abs/2208.04309) (2022).
8. Chu, Y. et al. Improving representation of high-frequency components for medical visual foundation models. *IEEE Transactions on Medical Imaging* (2025).
9. Zhang, S. et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint [arXiv:2303.00915](https://arxiv.org/abs/2303.00915) (2023).
10. Eslami, S., Meinel, C. & De Melo, G. Pubmedclip: How much does clip benefit visual question answering in the medical domain?. In *Findings of the Association for Computational Linguistics: EACL 2023*, 1181–1193 (2023).
11. Taha, A. A. & Hanbury, A. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging* **15**, 29 (2015).
12. Banerjee, S. & Lavie, A. 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*, 65–72 (2005).
13. Lau, J. J., Gayen, S., Ben Abacha, A. & Demner-Fushman, D. A. dataset of clinically generated visual questions and answers about radiology images. *Scientific data* **5**, 1–10 (2018).
14. Karim, R., Zhao, H., Wildes, R. P. & Siam, M. Med-vt: Multiscale encoder-decoder video transformer with application to object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6323–6333 (2023).
15. Pino, P., Parra, D., Besa, C. & Lagos, C. Clinically correct report generation from chest x-rays using templates. In *International Workshop on Machine Learning in Medical Imaging*, 654–663 (Springer, 2021).
16. Chen, Y.-C. et al. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120 (Springer, 2020).
17. Tian, D., Jiang, S., Zhang, L., Lu, X. & Xu, Y. The role of large language models in medical image processing: a narrative review. *Quantitative Imaging in Medicine and Surgery* **14**, 1108 (2023).
18. Szabó, A. & Laein, G. D. Comparative evaluation of large language models performance in medical education using urinary system histology assessment. *Scientific Reports* **15**, 31933 (2025).
19. Teng, X., Huang, L., Shen, Z. & Li, W. Improving intelligent perception and decision optimization of pedestrian crossing scenarios in autonomous driving environments through large visual language models. *Scientific Reports* **15**, 31283 (2025).
20. Dastani, M., Mardaneh, J. & Rostamian, M. Large language models' capabilities in responding to tuberculosis medical questions: testing chatgpt, gemini, and copilot. *Scientific Reports* **15**, 18004 (2025).
21. Li, C. et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* **36**, 28541–28564 (2023).
22. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nature Medicine* **31**, 943–950 (2025).
23. Xie, W., Wu, C., Zhang, X., Zhang, Y. & Wang, Y. Towards generalist foundation model for radiology. arXiv preprint [arXiv:2307.12680](https://arxiv.org/abs/2307.12680) (2023).
24. Li, W., Jia, F. & Hu, Q. Automatic segmentation of liver tumor in ct images with deep convolutional neural networks. *Journal of Computer and Communications* **3**, 146–151 (2015).
25. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**, 203–211 (2021).
26. Taleb, A. et al. 3d self-supervised methods for medical imaging. *Advances in neural information processing systems* **33**, 18158–18172 (2020).
27. Umirzakova, S., Muksimova, S., Baltayev, J. & Cho, Y. I. Force map-enhanced segmentation of a lightweight model for the early detection of cervical cancer. *Diagnostics* **15**, 513 (2025).
28. Ilesanmi, A. E., Ilesanmi, T. O. & Ajayi, B. O. Reviewing 3d convolutional neural network approaches for medical image segmentation. *Heliyon* **10** (2024).
29. Koukiou, G. & Anastasopoulos, V. Quantum 3d fft in tomography. *Applied Sciences* **13**, 4009 (2023).
30. Lu, Y., Liang, Y., Zhu, L. & Yang, Y. Freelong: Training-free long video generation with spectralblend temporal attention. *Advances in Neural Information Processing Systems* **37**, 131434–131455 (2024).
31. Nguyen, H., Ho, N. & Rinaldo, A. Sigmoid gating is more sample efficient than softmax gating in mixture of experts. *Advances in Neural Information Processing Systems* **37**, 118357–118388 (2024).
32. Wu, Z., Wei, J., Wang, J. & Li, R. Slice imputation: Multiple intermediate slices interpolation for anisotropic 3d medical image segmentation. *Computers in biology and medicine* **147**, 105667 (2022).
33. Vaswani, A. et al. Attention is all you need. *Advances in neural information processing systems* **30** (2017).
34. Heidari, A. A., Faris, H., Mirjalili, S., Aljarah, I. & Mafarja, M. Ant lion optimizer: theory, literature review, and application in multi-layer perceptron neural networks. *Nature-inspired optimizers: theories, literature reviews and applications* 23–46 (2019).
35. Naik, N. V. et al. Enhanced liver cancer detection: Leveraging pre-processing techniques for 3d ct volumes. *Gastroenterology & Endoscopy* (2025).
36. Alibrahim, H. & Ludwig, S. A. Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization. In *2021 IEEE congress on evolutionary computation (CEC)*, 1551–1559 (IEEE, 2021).
37. Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint [arXiv:2307.09288](https://arxiv.org/abs/2307.09288) (2023).
38. Loshchilov, I. & Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint [arXiv:1608.03983](https://arxiv.org/abs/1608.03983) (2016).
39. Woodworth, B. et al. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, 10334–10343 (PMLR, 2020).
40. Verma, K. & Maiti, A. Sine and cosine based learning rate for gradient descent method. *Applied Intelligence* **55**, 352 (2025).
41. Sun, Z. et al. Recent advances in lora: A comprehensive survey. *ACM Transactions on Sensor Networks* **18**, 1–44 (2022).
42. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**, 1929–1958 (2014).
43. Kobayashi, S., Akram, Y. & Von Oswald, J. Weight decay induces low-rank attention layers. *Advances in Neural Information Processing Systems* **37**, 4481–4510 (2024).
44. Li, C. et al. DeepSpeed data efficiency: Improving deep learning model quality and training efficiency via efficient data sampling and routing. In *Proceedings of the AAAI Conference on Artificial Intelligence* **38**, 18490–18498 (2024).
45. Hunhold, L. & Quinlan, J. Evaluation of bfloat16, posit, and takum arithmetics in sparse linear solvers. In *2025 IEEE 32nd Symposium on Computer Arithmetic (ARITH)*, 61–68 (IEEE, 2025).

46. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318 (2002).
47. Rouge, L. C.-Y. A package for automatic evaluation of summaries. In *InText Summarization brancheSout ProceedingSafthe ACL-04 WorkShop Wol &* (2004).
48. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. arXiv preprint [arXiv:1904.09675](https://arxiv.org/abs/1904.09675) (2019).
49. Gholamalizhad, H. & Khosravi, H. Pooling methods in deep neural networks, a review. arXiv preprint [arXiv:2009.07485](https://arxiv.org/abs/2009.07485) (2020).
50. Menze, B. H. et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging* **34**, 1993–2024 (2014).
51. Kovacs, B. et al. Data-centric strategies for overcoming pet/ct heterogeneity: Insights from the autopet iii lesion segmentation challenge. arXiv preprint [arXiv:2409.10120](https://arxiv.org/abs/2409.10120) (2024).
52. Artstein, R. Inter-annotator agreement. In *Handbook of linguistic annotation*, 297–313 (Springer, 2017).
53. Radford, A. et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PmLR, 2021).

Author contributions

Y. Lian designed the methodology, performed the validation and analysis, and wrote the original draft. Y. Xie conceptualized the study, supervised the project, and reviewed and edited the manuscript. Y. Jiang provided resources and was responsible for data curation. H.Y. conceptualized the study, supervised the project, was responsible for project administration and funding acquisition, and reviewed and edited the manuscript. All authors have read and approved the final manuscript.

Funding

This work was supported by the Major Science and Technology Special Project of Liaoning Province, China (Grant No. 2023020703-JH26/101-03, Research and Development of Multi-scenario Intelligent Robot Swarm Collaborative Command and Dispatch System, from December 2023 to December 2025).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.X. or H.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026