BIMCV-R: A Landmark Dataset for 3D CT Text-Image Retrieval Supplementary Materials

Yinda Chen[1,2], Che Liu[3], Xiaoyu Liu[1], Rossella Arcucci[3], and Zhiwei Xiong[1,2(✉)]

[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
[2]Anhui Province Key Laboratory of Biomedical Imaging and Intelligent Processing, Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3]Data Science Institute, Imperial College London
cyd0806@mail.ustc.edu.cn, zwxiong@ustc.edu.cn

1 Visual Results

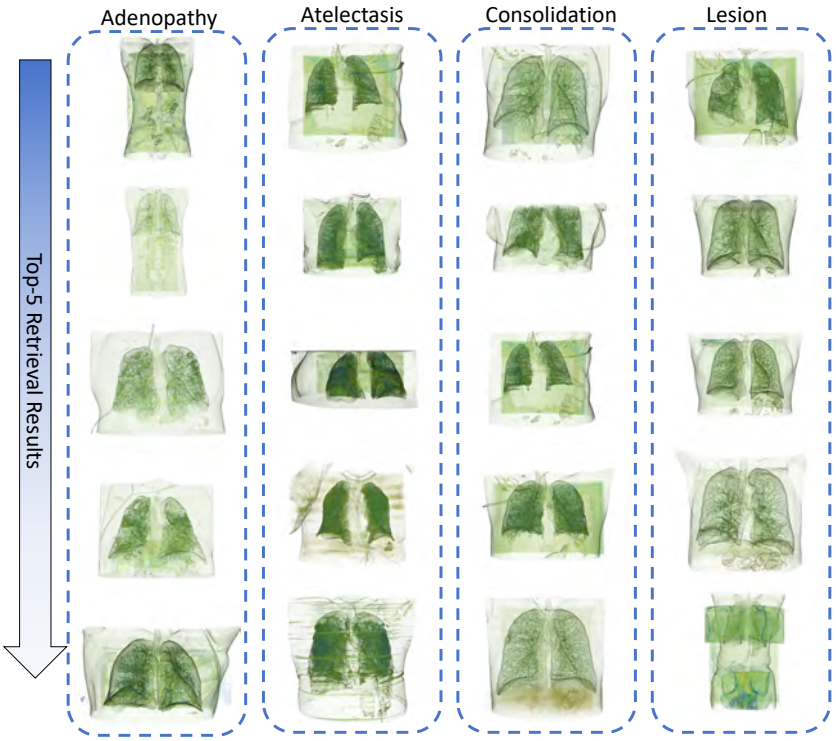

Fig. 1: Results of keyword-based retrieval from the BIMCV-R dataset, identifying corresponding CT volumes based on specified keywords. The illustration showcases the top 5 ranked samples.

2 Implementation Details

Table 1: Detailed Implementation Details for MedFinder Model.

Category	Details
Computing Resources	8 NVIDIA A40 GPUs
Batch Size	96
Number of Epochs	400
Learning Rate	1×10^{-6}
Alpha (α)	1
Text Encoder	BiomedCLIP (frozen)
Visual Encoder	3D ViT/Resnet (activated)
Data Augmentation	Noise addition, rotation, cutmix
Optimizer	SGD
Momentum	0.9
Weight Decay	1×10^{-5}
Learning Rate Decay	Cosine Annealing

3 Symbol Description

Table 2: Symbol Description.

Symbol	Meaning
R@K	Recall at K, a metric in retrieval tasks indicating the proportion in the top K results.
MdR	Median Rank, a metric in retrieval tasks indicating the average rank of the correct answer.
MnR	Mean Rank, a metric in retrieval tasks indicating the average rank of the correct answer.
L_{mse}	Mean Squared Error, a loss function used to measure the consistency of feature representations.
α	A weight coefficient in the total loss function balancing the MSE loss and matching loss.
T	The text description of a medical report.
T'	The sampled text representation.
F_{3D}	A 3D image encoder used to extract feature representations from 3D medical images.
Z	The feature representation of a 3D image.
Z_{cls}	Class-discriminative feature vector.
Z_{fusion}	A fused feature vector obtained through a cross-attention mechanism.
S	A similarity metric used to compute the similarity score between text and image features.
Score	The similarity score between text features and visual features.
L_{sim}	Similarity matching loss function, used to train the model to increase the similarity score.
L_{total}	Total loss function, combining MSE loss and matching loss, weighted by α.
P@K	Precision at K, a metric in retrieval tasks indicating the proportion in the top K results.