a). Contrastive Training

3D CT Volumes

Vision Encoder (DCFormer)

Frozen

Trainable

Matrix Similarity

	T_0	T_1	T_2	T_3	T_4
I_0	0.8	0.2	0.3	0.4	0.1
I_1	0.5	0.9	0.6	0.2	0.3
I_2	0.1	0.2	0.9	0.4	0.3
I_3	0.4	0.5	0.3	0.7	0.1
I_4	0.2	0.6	0.4	0.2	0.8

Text Encoder (ClinicalBert)

An 85×55 mm large solid mass…

Liver shows the following: cirrhotic changes…

Right lobe of liver shows an infiltrating retractile…

Radiology Reports

b). VLM Training

VLM Answers

Output Text tokens

LLM with LoRA

Image tokens

Text tokens

Multimodal Projector (MLP-Mixer)

Text Tokenizer

Vision Encoder (DCFormer)

Image Input

Generate a medical report based on this image.

Text input