Attention Consistent Longitudinal Medical Visual Question Answering Guided by Vision Foundation Models

Supplementary Material

1. Model parameter and memory usage

Table 1. Model Parameter Statistics

Component	Total	Trainable
Affine Registration	18.89K	18.89K
DINO Mask Generator	86.58M	0
Image Encoder	87.93M	87.93M
Adaptive Mask Generator	139.78K	139.78K
Symmetric Mask Fusion	4	4
Image Projector	10.67M	10.67M
Text Encoder	127.07M	127.07M
Generative Decoder	119.23M	119.23M
Mask Reconstruction Prediction Head	4.20M	4.20M
Total	**435.85M**	**349.27M**
Trainable: 80.14% (349,266,328 / 435,846,808)		

According to Table 1, the text encoder consumes much memory. However, given the general brevity of answers in longitudinal VQA tasks, future research may consider further reducing or even eliminating the text encoder.

2. More visualization examples of masks

To further demonstrate the model's interpretability and its performance in lesion difference identification, we provide 4 additional example masks. Darker shades indicate higher mask values. As shown in Figures 1 to 4, even when there is a degree of image offset between the two visits, the model remains capable of identifying the corresponding changes in the potential lesions, demonstrating robustness.

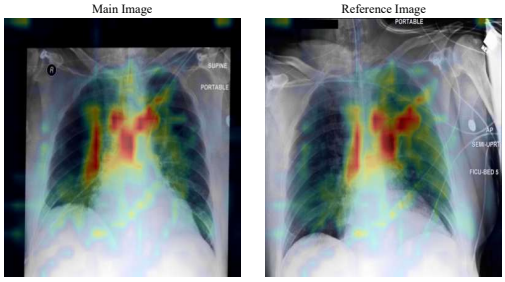

Figure 1. Mask visualization for sample 2614.

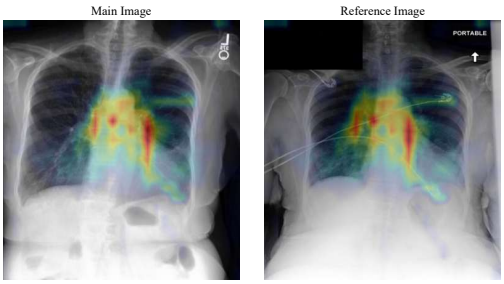

Figure 2. Mask visualization for sample 3039.

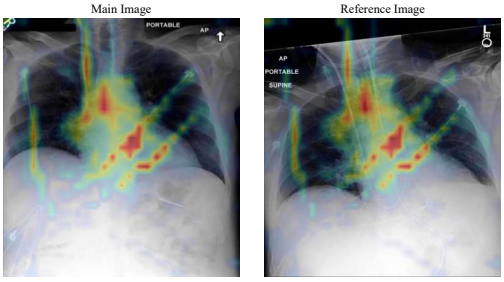

Figure 3. Mask visualization for sample 5094.

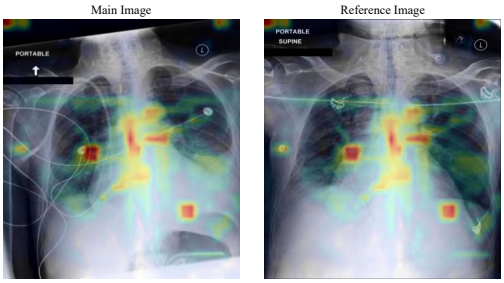

Figure 4. Mask visualization for sample 7467.