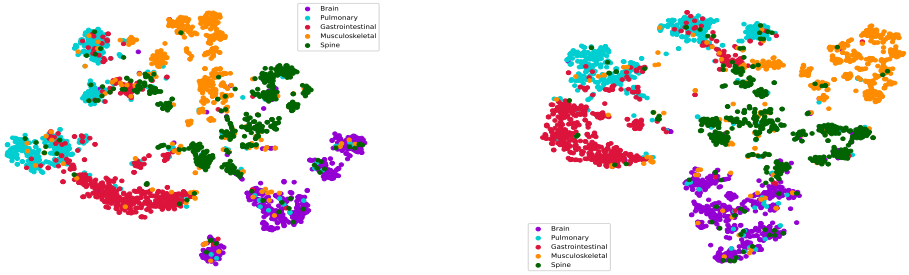
MedCLIP-SAM: Bridging Text and Images for Universal Medical Segmentation

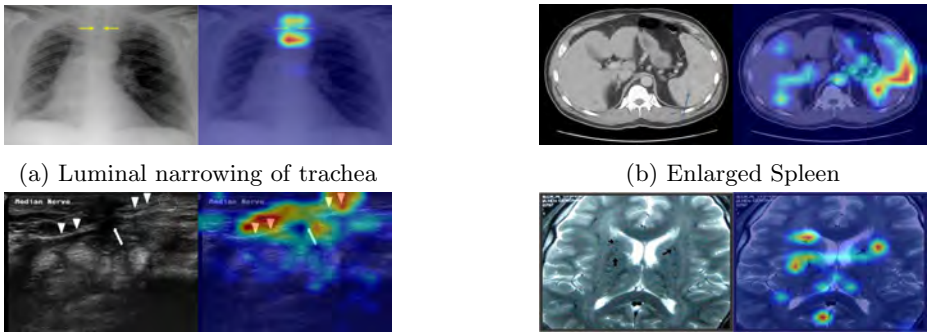
Taha Koleilat[1⋆], Hojat Asgariandehkordi[1], Hassan Rivaz[1], and Yiming Xiao[2]

[1] Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada
{taha.koleilat,hojat.asgariandehkordi,hassan.rivaz}@concordia.ca
[2] Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada
yiming.xiao@concordia.ca



(a) BiomedCLIP Pre-trained (b) Ours

Fig. 1: Visualization of 2D t-SNE image embeddings based on the anatomical categories of the medical diagnosis, computed based on the MedPix dataset, which was used to finetune the BiomedCLIP model.



(a) Luminal narrowing of trachea (b) Enlarged Spleen

(c) Scar tissue region (d) Foci of calcification

Fig. 2: Sample attention heatmaps with gScoreCAM after providing the text prompt to localize the region of abnormality in the radiology scans taken from the ROCO validation set. Arrows show the anatomical/pathological regions.

⋆ Corresponding author.