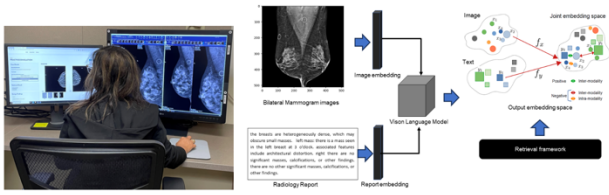
Knowledge-grounded Adaptation Strategy for Vision-language Models: Building Unique Case-set for Screening Mammograms for Residents Training

Aisha Urooj Khan[1], John Garrett[2], Tyler Bradshaw[2], Lonie Salkowski[2], Jiwoong Jason Jeong[3], Amara Tariq[1], and Imon Banerjee[1,3]

[1] Department of Radiology, Mayo Clinic
[2] Department of Radiology, UW Madison School of Medicine and Public Health
[3] School of Computing and Augmented Intelligence, Arizona State University



(a) (b)

Fig. 1: Multimodal learning for screening mammogram: (a) a session with radiology resident for the case review; (b) framework generating joint embedding space for bilateral mammogram and free-text radiology reports. Illustration of joint embedding space (right) is adapted from CrossCLR [22].

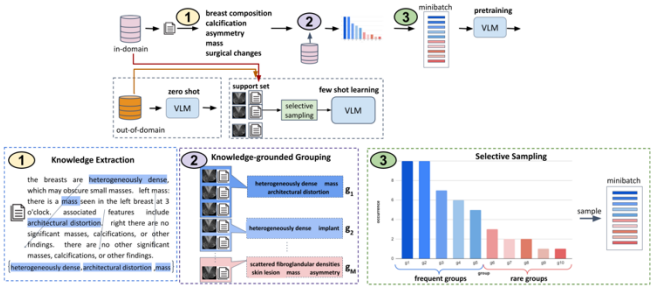
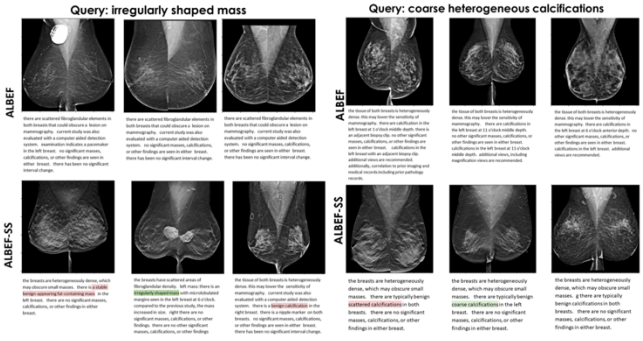


Fig. 2: Workflow for adapting the VLM with the proposed selective sampling to learn joint representation aware of fine-grained knowledge. The pretrained model is tested on out-of-domain data for zero shot evaluation. For few shot learning, support set is obtained from the training data to fine-tune model.



(a) (b)

Fig. 3: Qualitative results for Retrieval model. An example with highlighted green words is marked relevant by the radiologist for case build. Concepts highlighted with the pink show not exact but related findings in the image-report pair.