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

Aisha Urooj Khan[1], John Garrett[2], Tyler Bradshaw[2], Lonie Salkowski[2], Jiwoong 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

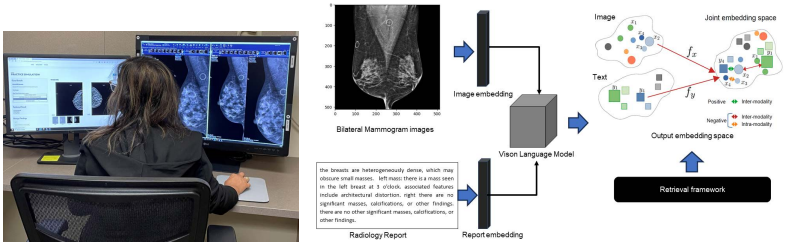
Abstract. A visual-language model (VLM) pre-trained on natural images and text pairs poses a significant barrier when applied to medical contexts due to domain shift. Yet, adapting or fine-tuning these VLMs for medical use presents considerable hurdles, including domain misalignment, limited access to extensive datasets, and high-class imbalances. Hence, there is a pressing need for strategies to effectively adapt these VLMs to the medical domain, as such adaptations would prove immensely valuable in healthcare applications. In this study, we propose a framework designed to adeptly tailor VLMs to the medical domain, employing selective sampling and hard-negative mining techniques for enhanced performance in retrieval tasks. We validate the efficacy of our proposed approach by implementing it across two distinct VLMs: the in-domain VLM (MedCLIP) and out-of-domain VLMs (ALBEF). We assess the performance of these models both in their original off-the-shelf state and after undergoing our proposed training strategies, using two extensive datasets containing mammograms and their corresponding reports. Our evaluation spans zero-shot, few-shot, and supervised scenarios. Through our approach, we observe a notable enhancement in Recall@K performance for the image-text retrieval task [4].

Keywords: multimodal understanding · retrieval · vision and language · mammogram

1 Introduction

According to the American Cancer Society (ACS) screening guidelines, women between 40 and 44 have the option to start screening with a mammogram every year and women 45 to 54 should get mammograms every year. This resulted in a huge number of screening mammogram exams at each healthcare institution and consumes significant radiologists' time for reading. During 12 weeks of required

[4] Code will be available at https://github.com/aurooj/VLM_SS.git



(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 [21].

residency training in breast imaging, the Accreditation Council for Graduate Medical Education (ACGME) requires residents to document a minimum of 300 interpretations of breast imaging exams (mammograms, ultrasounds, MRI) and there are no particular criteria for training case-selection [4]. Even after this requirement, the majority (59%) of residents do not feel prepared to read mammograms after completing their training [2,3]. Unfortunately, the number of fellowship-trained breast imaging radiologists is expected to decline and thus the majority of residents will face reading mammography as part of their eventual clinical practice. The fundamental fear of misdiagnosis (missing a cancer) and the feeling that residency does not fully prepare them to read mammograms, likely contributes to an increase in additional mammogram scans to confirm diagnosis and incur avoidable cost and effort [12]. Thus, providing adequate training with relevant case selection within radiology residency will benefit more women and bestow safer mammographic interpretation. However, hand-picking a set of such relevant cases is both time-consuming and challenging, as well as can introduce sampling bias and is unlikely to match the desired distribution. Furthermore, most PACS systems have search tools with very limited search criteria which often result in countless useless cases. Deep learning retrieval framework has the potential to automate and optimize case selection from 100,000's of cases based on multimodal data - imaging features and textual findings documented within the reports.

We develop a multimodal framework to automatize the relevant case-selection based on both text and image representation of the individual screening exams (Fig 1). However, there are inherent technical challenges for training such a model - (i) natural image pre-trained VLM is often unable to capture the radiology vocabulary with selective terms, and also natural image features do not correspond well with gray-scale and small mammography findings; (ii) relevant abnormal imaging findings (mass, calcification, architectural distortion, solitary dilated duct) are rare in screening mammogram which makes the model primarily learn the negative cases and omit the actual findings; (iii) syntactic difference between the semi-structured reports are minimal, and thus the reports with very different findings resulted similar embeddings; (iv) variations in breast density is often the most prominent image feature in mammogram and high density can occlude abnormal imaging features. To deal with the above-mentioned

challenges, we propose a *knowledge-based grouping of the mammogram cases, selective sampling, and hard-negative mining techniques for VLM model training.* We validate the efficacy of our proposed approach across two distinct VLMs: the in-domain VLM (MedCLIP) and the out-of-domain VLM (ALBEF). Our evaluation spans zero-shot, few-shot, and supervised scenarios using Institute X datasets containing mammograms and their corresponding reports. The model was also externally validated on screening mammogram data from Institute Y.

2 Related Work

i. Vision-language model in radiology - Several automated VLM efforts exist to generate radiology reports from images either as the template report generation task by filling with classified disease tag [18] or image-text generation task [14,16,1,13]. However, most of the current VLM models in radiology are focused on 2D chest X-rays due to the availability of open-source datasets [9,15]. Given the complexity of processing mammogram images(large dimension, varying density, multi-view), VLM literature is limited in the mammogram domain. *ii. Multi-modal Retrieval in radiology* - Recently, multimodal retrieval using image-text contrastive pre-training is gaining interest. For example, X-REM [8], CXR-RePaiR[5], ConVIRT [19], GLoRIA [6], and MedClip [17], leverage image-text contrastive pre-training to retrieve relevant radiology reports based on image and text embeddings. Despite these innovations, current frameworks face notable challenges: they lack strategies to preserve representation for rare cases crucial for embedding space integrity and struggle with mining 'hard-negatives' in radiology, particularly evident in mammogram studies where templated reports often inadequately describe distinct image features. Addressing these limitations is critical for enhancing the effectiveness of multimodal retrieval systems in medical imaging.

3 Methodology

Given a vision-language model $f(\theta)$, we want to train $f(\theta)$ effectively such that similar image-text pairs (I_p, T_p) are close to each other in semantic space. Negative pairs are often picked within a batch from a different data sample. For any given medical sub-domain, the vocabulary to describe the observations largely stays consistent, particularly in mammograms as the reports are formulated following the standardized BIRADS vocabulary [10] generated by the American College of Radiology (ACR). These image-report pairs can be grouped based on the important findings in a way that each image-report pair with the same concepts belongs to one group. Additionally, for mammograms, broad features are visually similar to each other and need a domain expert, i.e., a radiologist to examine for anomalies. Given the textual and visual similarity between the cases, there is a high chance that the sampled 'negative' image I_n or text report T_n has similar findings as the true pair does. This leads to confusion during model training because it might be pushing away semantically similar image-text pairs.

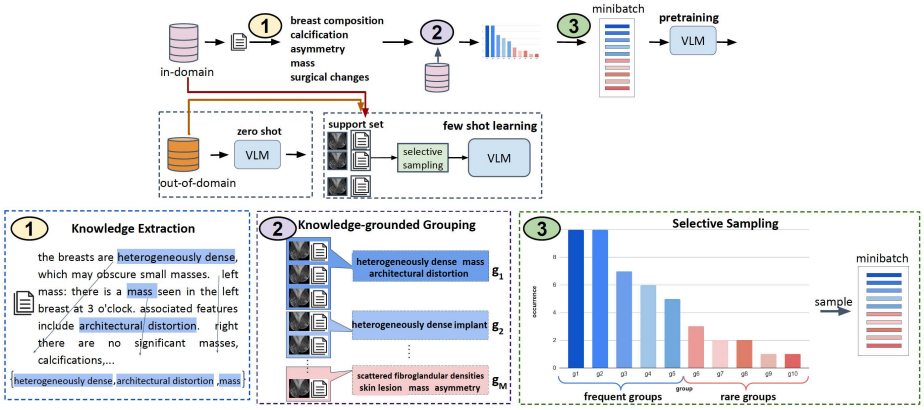


Fig. 2: Workflow for adapting the VLM with the proposed selective sampling to learn joint representation aware of fine-grained knowledge. The pretrained model is tested on out-of-domain data for zero-shot evaluation. For few-shot learning, a support set is obtained from the training data to fine-tune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

We propose a knowledge-grounded mini-batch sampling ensuring batch negatives to be coming from true negatives and minority cases are equally represented during training. This is achieved in three steps as described below:

1)Knowledge extraction: To form the groups, we leveraged the standard 54 unique BIRADS image descriptors and extracted the positive mentioned from the radiology reports which are lower cased and cleaned before extracting key concepts. For example, for the following text report:" *the breasts are heterogeneously dense, which may obscure small masses. left mass: there is a mass seen in the left breast at 3 o'clock. associated features include architectural distortion. right there are no significant masses, calcifications, or other findings*", the extracted group is {heterogeneously dense, mass, architectural distortion} based on the key concepts highlighted in blue. The abnormal image descriptors are primarily categorized into 5 groups - breast composition, calcification, asymmetry, mass, and surgical changes. All of these concepts except tissue density may or may not be present in the normal image without anomaly. We excluded all the negative and uncertain findings.

2) Knowledge grounded grouping: The presence of a key concept combination in any exam is considered a group such that every other image with the same key concepts present belongs to the same group. All text reports with the same key concepts (even ordered differently - $< A, B, C >$ vs $< B, C, A >$) belong to the same group. This yields a unique set of groups from the extracted knowledge for the given dataset. Formally, a group $g_i \in G^M$ for $i \in 1, 2, ..., M$ is a set of key concepts within an image extracted from the paired radiology report, where G^M is the set of M total groups extracted from the text reports. Negatives are defined as image-text pairs belonging to a different group while

some features may be common between them, e.g., positive group $< A, B >$ vs negative group $< A, B, C >$.

3) Selective Sampling: Given an image I_p and paired text report T_p as (I_p, T_p), a negative pair is denoted by (I_p, T_n) or (I_n, T_p), where I_n and T_n belong to an instance from a different group. For each pair (I_{p_i}, T_{p_i}) from group g_i, a negative image I_{n_j} or text T_{n_j} can be selected from group $g_j \in G^M$ when $j \neq i$. This approach while addressing the challenge of alike image-text pairs within a mini-batch, still faces the long-tail distribution challenge due to class imbalance. As frequent groups have a high chance of being sampled, rare groups often might never be seen during training. To address this problem, a mini-batch is sampled based on the group frequency. We define a heuristic-based boundary b to sample rare groups such that $b < batch\_size$ and $batch\_size - b$ instances are selected from groups with high occurrence, i.e., frequent groups. This ensures that b instances are coming from rare groups, where rare and frequent groups are empirically chosen based on the data distribution.

4. VLM Training: The proposed sampling strategy is used to sample mini-batches to train the vision-language model for contrastive learning. We use sampling strategy in two settings: pretraining and few-shot learning across two existing VLMs: ALBEF [11] and MedCLIP [17]. To measure the performance, we consider the Recall@K metric and report top-1, top-5, and top-10 performance. We consider it a success if any report with the same findings (hence the same group) appears in the top-K ranks.

4 Experiments and Results

Datasets: *Internal Dataset:* Using IRB approval, we collected 72,328 bilateral screening mammogram exams from 46,848 patients acquired between January 2016 and December 2018 from UW Madison health affiliated centers as our internal dataset. We randomly split the dataset into train-val-test with 70,238 $< image - report >$ pairs used for training, 1000 image-report pairs for validation, and 1000 image-report pairs as a test set respectively. We use a binary mask of thresholded pixel values to identify the largest connected component to crop the breast tissue area. The cropped R-MLO and L-MLO images are concatenated, zero-padded for maintaining the aspect ratio, and resized to 512×512 pixels. Reports are cleaned by lowercasing, punctuation removal, and extra spacing removal. The text is then split into sentences, each examined for key concepts: density, calcifications, asymmetry, architectural distortion, mass, and additional features. This grouping allows selective sampling during model training as described in 3. We find 1005 unique groups in the train set. Detailed group distribution is provided in the supplementary document.

External Dataset: With the Mayo Clinic IRB approval, the screening mammogram collected between 2018 - 2022 is used for external validation of our approach for supervised training as well as few-shot learning. The Mayo dataset has 8,172 training image-report pairs and 1,015 pairs in the test set. The test set is then used for external validation. The test set has 79 unique groups after preprocessing as described in section 3.

Task	Model	Internel test set			Externel test set		
		R@1	R@5	R@10	R@1	R@5	R@10
Image-to-Report	NN(k=10)	10.1	-	-	3.34	-	-
	ALBEF-Ret	12.9	37.0	47.2	19.00	50.21	65.76
	ALBEF-SS-PT (ours)	9.0	32.3	40.2	20.25	48.75	51.56
	ALBEF-SS-Ret (ours)	**30.5**	**53.9**	**61.3**	**21.61**	46.03	55.22
	MedCLIP	6.4	11.2	15.1	16.6	30.27	35.17
	MedCLIP-SS (ours)	5.10	10.60	14.90	4.28	11.69	20.98
Report-to-Image	NN(k=10)	26.4	-	-	36.95	-	-
	ALBEF-Ret	28.6	60.5	65.2	34.13	**82.98**	83.82
	ALBEF-SS-PT (ours)	19.4	60.7	67.6	**63.88**	81.73	84.76
	ALBEF-SS-Ret (ours)	**35.8**	**63.3**	**73.4**	54.70	81.94	**85.49**
	MedCLIP	26.70	48.40	56.30	0.31	20.77	22.02
	MedCLIP-SS (ours)	**31.5**	**62.3**	**66.2**	**0.52**	**21.4**	**24.22**

Table 1: Comparative retrieval results for the proposed knowledge grounded selective sampling (SS) on both internal (UW Madison) and external (Mayo Clinic) test sets. 'Ret':fine-tune models, 'PT':pre-trained model. Numbers are in percentages.

Implementation Details: ALBEF [11] is a VLM with image-text contrastive loss. We pre-train ALBEF on UW Madison image-report pairs, followed by a retrieval-only task Image Text Matching (ITM) for fine-tuning the pretrained backbone named ALBEF-Ret. For a 512×512 image and the patch size of 16×16, image encoder takes 1024 patch tokens in the ALBEF model. We train ALBEF with (ALBEF-SS) and without (ALBEF-Ret) the proposed selective sampling. We evaluate MedCLIP [17] pretrained on CheXpert dataset [7] and MIMIC-CXR [9] for zero-shot, initialize model weights for few-shot learning, and train MedCLIP on the 2D mammogram images for fully supervised backbone. Similar to ALBEF, we also trained MedCLIP with (MedCLIP-SS) and without (MedCLIP) the proposed selective sampling. For full training, we consider the top 20 groups w.r.t the number of samples as frequent groups out of a total of 1005 unique groups. We use batch size=8 and boundary b=3 for random sampling of frequent and rare groups, i.e., for R=0.375 - 5 instances belong to frequent groups, and 3 are sampled from the set of rare groups. All training parameters except the hyperparameters stayed the same across models.

Results: We evaluate the learned joint embedding using image↔text retrieval (ITR) as our downstream task. We compare ALBEF with ALBEF-SS, and MedCLIP with MedCLIP-SS to assess the impact of selective sampling during training. We observe improvement for both VLMs with selective sampling for image-to-report and report-to-image retrieval on our internal test set as well as external test data. Table 1 presents the complete results on the internal and external data. More specifically, on the internal test set, ALBEF-SS-Ret obtains 17.6% ↑ gain in R@1 performance, ∼17%↑ improvement in R@5, and 14.1%↑ increase in R@10 score over ALBEF-Ret model for image-to-report retrieval. For report-to-image retrieval, ALBEF-SS-Ret improves by 7.2% ↑ at R@1, 2.8% ↑ at R@5, and 8.2% ↑ at R@10 scores. MedCLIP-SS achieves comparable results to the MedCLIP baseline for R@5 and R@10. For report-to-image retrieval, MedCLIP-SS achieves a performance gain of 4.8%↑ in R@1, 1.8%↑ as R@5, and with a significant margin of ∼10%↑ in R@10 respectively. Overall, we observe that image-to-report retrieval is a more challenging task for VLMs compared to report-to-image retrieval. On the external test set, ALBEF-SS-Ret model al-

Task	K	Model	Internal test set			External test set		
			R@1	R@5	R@10	R@1	R@5	R@10
Image-to-Report	ZS	MedCLIP-ViT	1.9	12.0	20.5	**25.71**	38.42	40.79
		ALBEF-mscoco	16.8	32.0	40.5	14.61	36.01	43.11
		ALBEF-flickr30k	20.0	31.1	37.5	7.83	33.82	40.29
		ALBEF-SS-Ret (ours)	-	-	-	21.61	**46.03**	**55.22**
	10	MedCLIP	0.1	3.1	6.8	**32.36**	**48.43**	**57.09**
		MedCLIP-SS	**2.2**	**8.0**	**14.1**	18.00	36.22	41.44
		ALBEF	19.5	46.9	55.0	0.3	29.96	55.01
		ALBEF-SS-Ret	**25.40**	**48.10**	**57.40**	**20.88**	**46.76**	**56.47**
Report-to-Image	ZS	MedCLIP-ViT	24.1	42.6	46.6	35.66	55.37	81.48
		ALBEF-mscoco	5.6	41.2	48.7	1.36	35.07	68.37
		ALBEF-flickr30k	2.2	44.3	50.5	0.32	61.17	57.74
		ALBEF-SS-Ret (ours)	-	-	-	**54.70**	**81.94**	**85.49**
	10	MedCLIP	3.3	38.6	46.4	1.57	36.64	**57.20**
		MedCLIP-SS	**6.6**	33.2	**54.6**	**36.95**	**55.53**	56.68
		ALBEF	32.9	65.9	75.0	36.74	68.99	81.84
		ALBEF-SS-Ret	31.6	**67.3**	73.2	35.39	**78.29**	80.06

Table 2: Zero-shot (ZS) and few-shot (K=10) results for image↔report retrieval. MedCLIP-ViT is pretrained on chest x-rays [9], [7], MedCLIP and MedCLIP-SS are trained on the screening mammogram exams. Numbers are in percentages.

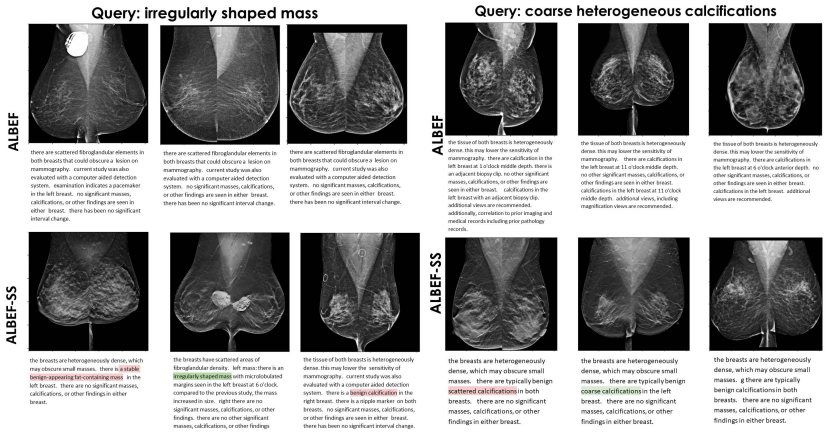
though improves over ALBEF by 2.61% in terms of R@1, performance is hurt on R@5 and R@10. Similar behavior is observed for MedCLIP-SS as well. However, we notice a consistently significant improvement in both ALBEF-SS-Ret and MedCLIP-SS for report-to-image retrieval. MedCLIP-SS consistently performs better than MedCLIP in terms of R@1, R@5, and R@10 respectively.

Zero-shot retrieval: We further compare the zero-shot performance on the external test set from Mayo using off-the-shelf models: MedCLIP-ViT, MSCOCO-pretrained ALBEF, and Flick30K-pretrained ALBEF and compare to ALBEF-SS-Ret pretrained on ∼70K internal samples. For image-to-report, MedCLIP-ViT obtains the best R@1 score: 25.7% vs. second-best 21.61% from ALBEF-SS-Ret. ALBEF-SS-Ret outperforms MedCLIP-ViT on R@5 and R@10 by 7.61% ↑ and 14.43% ↑ respectively. For report-to-image retrieval, ALBEF-SS-Ret outperforms MedCLIP-ViT by 19.04% ↑, 26.57% ↑, and 4.01% ↑ in terms of R@1, R@5, and R@10 respectively. See table 2 for complete results.

Few-shot retrieval: For the few-shot learning setup, we sampled up to K=10 instances for each group from an internal training set. For groups with less than 10 instances, we keep all available instances. This resulted in 3,331 unique training image-report pairs.

Internal test set: For image-to-report retrieval evaluation, ALBEF-SS-Ret outperforms ALBEF on all three metrics. MedCLIP-SS also demonstrates consistent improvements across all metrics with at least 50% relative performance gain over MedCLIP. For report-to-image, MedCLIP shows improvement in R@1 (3.3%↑) and R@10 (8.2%↑). ALBEF-SS-Ret shows overall comparable performance to ALBEF with a slight gain in the R@5 score.

External test set: We observe that ALBEF-SS-Ret performs significantly better than its counterpart (R@1 score: 20.88% vs 0.3%, R@10: 46.76% vs 29.96%) when doing image-to-report retrieval during external validation. For report-to-image retrieval, it improves R@5 by approx. 10 points while performing comparable to ALBEF on R@1 and R@10. MedCLIP-SS, in comparison with MedCLIP, also shows significant improvement for R@1 (36.95% vs 1.57%) and R@5 (55.53%



(a) (b)

Fig. 3: Qualitative results for Retrieval model. Samples with highlighted green words are marked relevant by a radiologist and in pink, show not exact but related findings in the image-report pair.

vs 36.64%) scores respectively on report-to-image retrieval task, but shows the opposite trend on image-to-report retrieval. Overall, we observe that selective sampling consistently benefits the ALBEF model for both internal and external validation. MedCLIP-SS, on the other hand, while being beneficial for internal testing as well as for external validation of report-to-image retrieval performance, seems to be less effective for out-of-domain image-to-report retrieval. This is consistent with the trends observed while performing external validation of MedCLIP-SS when trained on the full training set. We need to re-calibrate the frequent groups to benefit from selective sampling based on the support set's group distribution.

	Image-to-Report			Report-to-Image		
Method	R@1	R@5	R@10	R@1	R@5	R@10
(1) R=0.25	0.4	1.5	2.4	3.2	29.2	41.6
(2) R=0.38	0.4	2.8	**8.7**	15.7	30.7	51.3
(3) R=0.50	0.1	1.8	5.2	**17.7**	**41.2**	**58.9**
(4) R=0.75	**0.5**	5.2	7.7	1.4	26.3	28.9
(5) w/ B shuffle	0.3	1.8	6.8	17.1	24.7	42.6
(6) w/o B shuffle	**0.4**	**2.8**	**8.7**	**15.7**	**30.7**	**51.3**
(7) Freq. groups, fixed	17.00	44.30	55.30	32.90	66.50	73.80
(8) Freq. groups, recalibrate	**25.40**	**48.10**	**57.40**	31.60	**67.30**	73.20

Table 3: Ablations for the proposed sampling strategy on Institute X using MedCLIP-SS model. B=batch size, R=ratio of frequent groups to rare groups in a batch.

Ablations and Analyses: Table 3 reports the selected ablations from our detailed analyses regarding important hyperparameters such as #samples from frequent vs. rare groups, recalibrating no. of frequent groups with change in data distribution that happens during few-shot learning, and choice of mini-batch shuffling after our selective sampling. We used MedCLIP-SS with few-shot learning (K=10) in all ablations unless specified otherwise. See additional results in the supplementary document.

5 Discussion and Conclusion

Training a large network on medical data, particularly with contrastive loss, is always challenging when the dataset is highly influenced by the majority of 'normal' cases and instances with compelling representation (image or textual) are extremely rare. Our proposed knowledge-grounded selective sampling strategy helps the contrastive model training by ensuring the sampling of the true negatives and equalizing representation of rare cases. We observed improvement in the retrieval performance with the selective sampling strategy, especially for the ALBEF model. For MedCLIP, we observed improvement for internal evaluation; however, there was no improvement on the external dataset for image-to-report which could be based on the fact that image-to-text retrieval is a more challenging task and we didn't pre-train the MedCLIP on the mammogram dataset. However, we still observed MedCLIP performance improvement on the external dataset for report to image particularly in R@1 and R@5 for few-shot learning. On the zero-shot performance, our pre-trained model also outperformed all the baselines, including MedCLIP-VIT, on the external dataset for both image-to-report and report-to-image retrieval tasks. It is also highlighted in the domain of LLMs that few-shot learning can be highly sensitive to the quality of the demonstrations, emphasizing the need for strategies to strategically select few-shot [20].

Based on the ablation study, we also present the fact that proposed selective sampling can help to train the VLM model with a smaller batch size for a limited resource setting. However, thorough experimentation needs to be done with intelligent sampling to balance the groups for larger batch sizes to properly understand the relationship between the number of groups and the batch size.

In summary, our proposed sampling strategy lays the groundwork to rethink data sampling strategies for effective training of multimodal networks as well as for in-context learning, case-in-point, vision-language models grounded in the multimodal data for medical contexts.

Acknowledgement. Research reported in this paper was supported by NCI of the National Institutes of Health under award number 1R37CA262110-01A1 and NIH/NCI, U01 CA269264-01-1.

Disclosure of Interests. John Garrett has received research grants from Flywheel.io, GE Healthcare, and Optum, is a member of the SIIM Machine Learning Tools and Research Committee, owns stock in NVIDIA, and is an advisor to and holds equity in RadUnity.

References

1. Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., Fahmy, A.: Automated radiology report generation using conditioned transformers. Informatics in Medicine Unlocked **24**, 100557 (2021) 3
2. Bassett, L.W., Monsees, B.S., Smith, R.A., Wang, L., Hooshi, P., Farria, D.M., Sayre, J.W., Feig, S.A., Jackson, V.P.: Survey of radiology residents: breast imaging training and attitudes. Radiology **227**(3), 862–869 (2003) 2

3. Beam, C.A., Layde, P.M., Sullivan, D.C.: Variability in the interpretation of screening mammograms by us radiologists: findings from a national sample. Archives of internal medicine **156**(2), 209–213 (1996) 2
4. Davis, D.J., Ringsted, C.: Accreditation of undergraduate and graduate medical education: how do the standards contribute to quality? Advances in health sciences education **11**, 305–313 (2006) 2
5. Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., Rajpurkar, P.: Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In: Machine Learning for Health. pp. 209–219. PMLR (2021) 3
6. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021) 3
7. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019) 6, 7
8. Jeong, J., Tian, K., Li, A., Hartung, S., Adithan, S., Behzadi, F., Calle, J., Osayande, D., Pohlen, M., Rajpurkar, P.: Multimodal image-text matching improves retrieval-based chest x-ray report generation. In: Medical Imaging with Deep Learning. pp. 978–990. PMLR (2024) 3
9. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019) 3, 6, 7
10. Lazarus, E., Mainiero, M.B., Schepps, B., Koelliker, S.L., Livingston, L.S.: Bi-rads lexicon for us and mammography: interobserver variability and positive predictive value. Radiology **239**(2), 385–391 (2006) 3
11. Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S., Xiong, C., Hoi, S.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS (2021) 5, 6
12. Miglioretti, D.L., Gard, C.C., Carney, P.A., Onega, T.L., Buist, D.S., Sickles, E.A., Kerlikowske, K., Rosenberg, R.D., Yankaskas, B.C., Geller, B.M., et al.: When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology **253**(3), 632–640 (2009) 2
13. Mohsan, M.M., Akram, M.U., Rasool, G., Alghamdi, N.S., Baqai, M.A.A., Abbas, M.: Vision transformer and language model based radiology report generation. IEEE Access **11**, 1814–1824 (2022) 3
14. Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M.: Progressive transformer-based generation of radiology reports. arXiv preprint arXiv:2102.09777 (2021) 3
15. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017) 3
16. Wang, Y., Lin, Z., Xu, Z., Dong, H., Luo, J., Tian, J., Shi, Z., Huang, L., Zhang, Y., Fan, J., et al.: Trust it or not: Confidence-guided automatic radiology report generation. Neurocomputing p. 127374 (2024) 3
17. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text (2022) 3, 5, 6

18. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 72–82. Springer (2021) 3
19. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022) 3
20. Zhao, Z., Wallace, E., Feng, S., Klein, D., Singh, S.: Calibrate before use: Improving few-shot performance of language models. In: International Conference on Machine Learning. pp. 12697–12706. PMLR (2021) 9
21. Zolfaghari, M., Zhu, Y., Gehler, P., Brox, T.: Crossclr: Cross-modal contrastive learning for multi-modal video representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1450–1459 (2021) 2