Finetuning Medical Vision-Language Models for Interpretable and Fine-grained Disease Classification

Sofija Engelson[a], Jan Ehrhardt[a,b], and Heinz Handels[a,b]

[a]German Research Center for Artificial Intelligence, Maria-Goeppert-Straße 15, 23562 Lübeck, Germany
[b]Institute of Medical Informatics, University of Luebeck, Ratzeburger Allee 160, 23562 Lübeck, Germany

ABSTRACT

Medical vision-language models (VLMs) hold great promise for interpreting clinical data, yet struggle with fine-grained concept recognition. In this work, we propose a finetuning approach for medical VLMs that improves the discrimination of classes and class-descriptive concepts, thereby increasing interpretability by refining the reasoning behind each class prediction. To this end, we use adapter-layers for finetuning a pre-trained VLM, introduce prior knowledge about class-concept relations, and leverage supervised contrastive learning to learn inter-subject correspondences. We evaluate our method on two use cases and vision-language models: melanoma classification with MONET and multi-label lung disease classification with MedCLIP. Our results demonstrate that finetuning with as little as 5% of training data yields better results than zero-shot prediction of the original VLM, especially for fine-grained class definitions. Moreover, we show that our approach shows better performance in regard to AUC and F1-score than linear probing and prompt tuning, and it is more robust to class imbalance. This highlights the strength of leveraging prior knowledge in medical VLM finetuning for maintaining the generalization capabilities of foundation models while delivering competitive results on specialized tasks.

Keywords: Medical Vision-Language Models, CLIP, Finetuning, Concept-based Explanations.

1. INTRODUCTION

With the recent advancement in the field of computer vision as well as natural language processing, the modality alignment of image and text becomes increasingly relevant. In the context of contrastive learning-based vision-language models, Radford et al.[1] introduced CLIP (Contrastive Language-Image Pre-training) trained solely on unlabelled data with the aim to perform zero-shot image-text matching and retrieval. In a survey on CLIP models in medical imaging, Zhao et al.[2] identify three key challenges when it comes to the transfer of CLIP-based pre-training from natural images to medical images. First, CLIP was trained to align global information, that is, the whole image is matched with its corresponding text. In medical applications, pathologies are usually described by single keywords in the report referring to image subregions. Hence, CLIP models for medical data need to learn features at multiple scales. Second, medical datasets with paired images and reports are scarcely available. Comparing medical and general-domain VLMs reveals that the amount of training data available for the former is a multitude lower.[3] Third, medical semantics are highly specialized, complex, and carry hierarchical dependencies. Not accounting for these challenges can lead to degraded performance and hidden shortcut learning.[2] Due to the high stakes of biased decisions in healthcare, the interpretability of decision-support systems becomes particularly important, and building user trust in the AI system forms a prerequisite for deployment. In the area of interpretability, the trend is to move from local explanation methods to global, instance-independent explanations such as prototypes or concepts.[4,5] Hence, to improve interpretability, there is a need for medical VLMs to differentiate classes, but also accompanying class-descriptive concepts. A concept is a semantic characteristic that belongs to the class such as the location, the disease spread, or a visual feature attributed to the disease. In the following examples, the class is marked in bold and the class-descriptive concepts

Further author information: (Send correspondence to S.Engelson.)
S.Engelson: E-mail: sofija.engelson@dfki.de, Telephone: +49 451 3101 5638

Medical Imaging 2026: Computer-Aided Diagnosis, edited by Axel Wismüller, Thomas Martin Deserno, Proc. of SPIE Vol. 13926, 1392605 · 2026
Published by SPIE · 1605-7422 · doi: 10.1117/12.3086032

in italic: "*Moderate* **consolidation** at the *lower lung zone*." or "An image of a skin lesion with *diffuse regular pigmentation*, which indicates a **nevus**.". As illustrated in these examples, class-descriptive concepts help explain why a diagnosis is assigned and indicate the image regions associated with the underlying pathology, thereby enhancing interpretability and verifiability. Furthermore, class-descriptive concepts are naturally included in the description of medical images or in radiology reports and should therefore be represented within the VLM's text encoder. However, an initial experiment evaluating the zero-shot performance of MONET,[6] a foundation model trained on dermatological images, on the Derm7p[7] test dataset shows that for the given example the prediction performance on class-level can not be sustained, when the differentiation between classes becomes more fine-grained. Table 1 shows the differences between classes and concepts for pigmentation, one of the seven criteria defined in the Derm7p dataset,[7] and describes the mentioned experiment.

	criterion	**Concept**				
	pigmentation	Absent	Localized Regular	Diffuse Regular	Localized Irregular	Diffuse Irregular
Class	nevus	✓	✓	✓		
	melanoma				✓	✓

Table 1: Definition of classes and concepts for the criterion *pigmentation* within the seven point evaluation in Derm7p.[7] At class-level, a distinction is made only between pigmentations characterizing a nevus and pigmentations that are typical for melanoma. At concept-level, the pigmentation is described by five characteristics. The mean accuracy for zero-shot prediction of the expressions of the criterion pigmentation using MONET[6] drops heavily from 0.59 on class-level to 0.29 when class and concept need to be differentiated (class-concept-level).

In this study setup, an existing medical VLM and a labelled dataset for a specified task are given, and we aim to finetune the VLM such that the new knowledge of the specified task is induced into the VLM while keeping the general knowledge from original training. We measure the success of finetuning by examining the VLM's ability to distinguish fine-grained semantics of the class label. Existing research presents multiple methods for an adaptation of the network architecture,[8–11] but falls short in investigating further options for suitable adaptations of the loss function for finetuning. Especially, when finetuning on a small dataset, using the same loss function as for training the VLM, i.e., a training strategy that exploits unsupervised contrastive learning, creates a noisy supervision signal by separating all non-matching samples in a batch, regardless of semantics. A more granular class definition, as explored in this work, further enforces the problem of cases being falsely pushed apart during unsupervised contrastive learning.

1.1 Related Work

Many authors have identified the listed challenges of medical VLM training and addressed these in their work. To enforce the learning of multiscale features, some authors aim to progress the cross-modal alignment of local information, that is, an image subregion and single words.[12–16] More concretely, Huang et al.[12] and Dawidowicz[13] introduce a local contrastive loss on the attention-weighted features of image regions and features of single words. These approaches utilize the fine-grained intra-subject correspondences, but ignore the fact that with smaller amounts of available training data, many pairs in a batch are falsely pushed apart despite sharing high-level semantics. To address this issue, Liu et al.[17] calculate the similarity of reports via BioClinicalBERT and, in this way, classify into positive, neutral, and negative alignment of reports further used in the loss calculation. Wang et al.[3,18] combine synonym descriptions of classes (e.g., the disease) to encourage the model to learn inter-subject correspondences. Another strand of research infuses structured expert knowledge as described in Unified Medical Language System (UMLS)[19] into the network training, either by leveraging entity-relation triplets to improve image-text alignment[20] or by replacing the input texts with entity-relation triplets.[21,22]

Building on the related work for original training VLMs, we propose to leverage class-concept relations as prior knowledge to address the above-mentioned challenges. Our proposed method SCALE (**S**upervised **C**ontrastive **A**dapter **L**earning for fine-grained **E**valuation) combines CLIP adapter[8] as a finetuning technique, domain knowledge about hierarchical relationships on class- and concept-level, and supervised contrastive learning.[23,24] In this way, we aim to improve interpretability by distinguishing fine-grained expressions of classes, i.e.,

class-concept-combinations. We apply our method to two use cases and their medical domain-specific VLMs, i.e., MONET[6] for the dermatological use case, and MedCLIP[3] for lung and heart disease classification on chest X-rays. Furthermore, we assess the performance for varying amounts of labelled data and compare the results to an image-based supervised baseline and a prompt tuning approach.

2. METHODS

Given a pre-trained medical VLM and labels for classes and class-descriptive concepts for an unseen training dataset, we measure the success of finetuning by examining the VLM's ability to distinguish fine-grained semantics of the class label. To this end, the pairs of images I and texts T are passed through frozen image and text encoders. Only appended adapter layers, as proposed by Gao et al.,[8] are trained in order to finetune the VLM. Adapter layers enable the integration of new knowledge into the model, while mitigating the risk of overfitting and catastrophic forgetting. The proposed loss function includes supervised loss terms for classes, concepts, and class-concept combinations with the goal of capturing multiscale features and hierarchical semantics within medical data. Furthermore, supervised contrastive learning is employed in our training strategy to exploit inter-subject correspondences. When there is little training data available, not accounting for semantic correspondences in a batch can lead to semantically similar image-text pairs to be falsely pushed apart in contrastive learning. Fig. 1 shows an overview of our proposed finetuning pipeline.

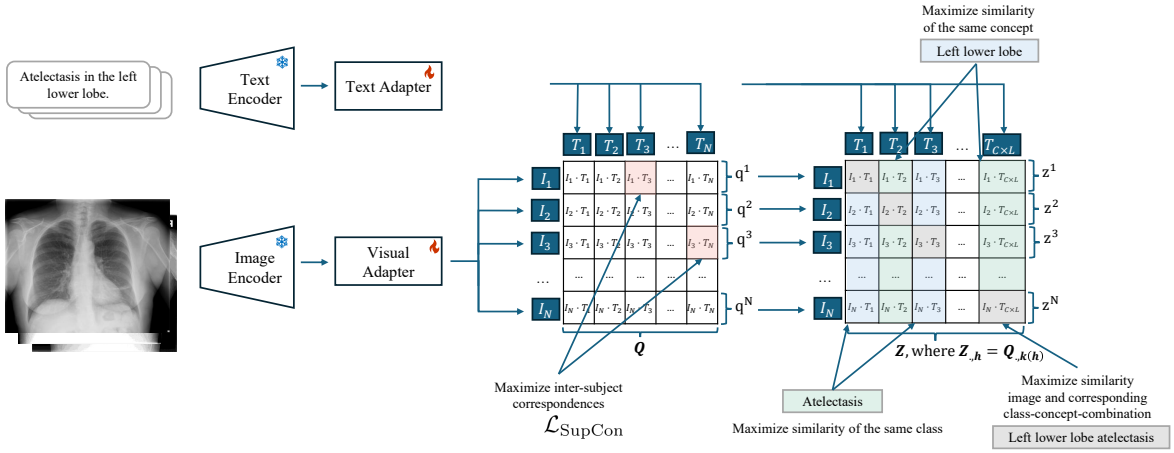


Figure 1: Overview of SCALE's training process. A batch of N image-text pairs is processed by the frozen visual and text encoders. Next, the resulting embeddings are passed through the respective adapter. The original embeddings and the adapted embeddings are blended and then used for loss calculation. The proposed loss function maximizes the similarity between same classes, same concepts, and corresponding class-concept-combinations in prediction matrix $\boldsymbol{Z}$. Furthermore, supervised contrastive learning maximizes the similarity of semantically similar image-text pairs across patients in a batch.

2.1 Network Architecture

Instead of updating the parameters of the whole VLM backbone, Gao et al.[8] propose to introduce adapter layers and blend the embeddings resulting from the frozen backbone with the embeddings resulting from the adapters. An adapter A is defined as follows:

$$A(f) = ReLU(f^T W_1)W_2, \tag{1}$$

where f are the visual features or text embeddings from the frozen CLIP backbone, W_1 and W_2 are weights of the bottleneck linear layers.[8] The residual-style blending happens according to $f' = \alpha A(f)^T + (1-\alpha)f$ where α is also a learnable parameter. For our analyses, we used two adapters for both encoding branches.

2.2 Loss Calculation

We assume that each concept $l \in L$ belongs to C_l, a set of one or more classes, and each class $c \in C$ is associated with a set of concepts, denoted by L_c. These class-concept relations are available as prior knowledge that has been assigned according to the 7-point checklist for skin lesion evaluation in the dermatological use case[7] and an aggregated version of locations according to UMLS[19] for the disease classification on thoracic X-rays[25] (see Sec. 2.3.1 for more details). Matrix $\boldsymbol{Q}$ contains all normalized cosine similarities between the image embeddings I and text embeddings T. For calculating the supervised loss terms, $\boldsymbol{Q}$ is transformed into $\boldsymbol{Z}$ such that $\boldsymbol{Z}_{.,h} = \boldsymbol{Q}_{.,k(h)}$ where $h \in \{1, 2, ..., C * L\}$ is the index of all possible class-concept combinations and k is the index of the first column associated with class-concept combination h in the columns of $\boldsymbol{Q}$. The prediction vector for image $i \in I$, where $I = \{1, 2, ..., N\}$, is denoted by $\boldsymbol{z}^i = \boldsymbol{Z}_{i,.}$, where $\boldsymbol{Z} \in \mathbb{R}^{N \times (C*L)}$. The concept and class prediction vectors $\tilde{\boldsymbol{z}}^i_{\text{concepts}}$ and $\tilde{\boldsymbol{z}}^i_{\text{classes}}$, respectively, are aggregated versions of $\boldsymbol{z}^i$ according to:

$$\tilde{\boldsymbol{z}}^i_{\text{concepts}} = (z_l)_{l=1}^{L}, \text{ where } \; z_l = \max\{\boldsymbol{z}^i_{C_l}\}$$
$$\tilde{\boldsymbol{z}}^i_{\text{classes}} = (z_c)_{c=1}^{C} \; , \text{ where } \; z_c = \max\{\boldsymbol{z}^i_{L_c}\}.$$

Let the ground truth $\boldsymbol{y}^i = (\boldsymbol{l}^i, \boldsymbol{c}^i, \boldsymbol{g}^i)$ for image i be given as a triple of one-hot encoded ground truth vectors for the concepts $\boldsymbol{l}^i$, the classes $\boldsymbol{c}^i$, and their combinations $\boldsymbol{g}^i = \boldsymbol{G}_{i,.} \in \{0, 1\}^{1 \times (C \cdot L)}$. The first $\mathcal{L}_{\text{CE}}$ term is the cross-entropy loss of the prediction vector $\boldsymbol{z}_i$ to the ground truth class-concept combinations $\boldsymbol{g}^i$. In order to increase the similarity of semantically similar concepts that belong to differing classes or similar classes that belong to differing concepts, we added two supervised loss terms accordingly. Furthermore, we included a supervised contrastive loss term $\mathcal{L}_{SupCon}$ to enable the model to learn inter-subject correspondences. For multi-label classification in the second use case, as proposed by Zhang et al.,[24] sample i is treated as a separate sample for each class-concept combination $\boldsymbol{G}_{i,h}$. Given $S(i) \equiv I \setminus \{i\}$, $\mathcal{L}_{SupCon}$ is defined as follows:

$$\mathcal{L}_{SupCon}(\boldsymbol{G}, \boldsymbol{Q}) = \sum_{i=1}^{N} \mathcal{L}_i^{SupCon}(\boldsymbol{g}^i, \boldsymbol{q}^i) = \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\boldsymbol{q}^i \cdot \boldsymbol{q}^p / \tau)}{\sum_{s \in S(i)} \exp(\boldsymbol{q}^i \cdot \boldsymbol{q}^s / \tau)} \tag{2}$$

where $\boldsymbol{q}_i = \boldsymbol{Q}_{i,.}$ and τ is a temperature parameter, that controls the sharpness of a distribution.[23] Given that sample i is the anchor, $P(i) \equiv \{p \in S(i) : \boldsymbol{g}^p = \boldsymbol{g}^i\}$ is the set of indices of all positives, i.e., all samples with the same label as anchor i. The cardinality of this set is denoted by $|P(i)|$. The total loss function is described by:

$$\mathcal{L} = \sum_{i \in N} \mathcal{L}_{\text{CE}}(\boldsymbol{g}^i, \boldsymbol{z}^i) + \lambda_1 \mathcal{L}_{\text{CE}}\big(\boldsymbol{l}^i, \tilde{\boldsymbol{z}}^i_{\text{concepts}}\big) + \lambda_2 \mathcal{L}_{\text{CE}}\big(\boldsymbol{c}^i, \tilde{\boldsymbol{z}}^i_{\text{classes}}\big) + \lambda_3 \mathcal{L}_{SupCon}(\boldsymbol{G}, \boldsymbol{Q}) + \lambda_4 \mathcal{L}_{SupCon}(\boldsymbol{G}, \boldsymbol{Q}^{\intercal}). \tag{3}$$

Hyperparameter λ_1 to λ_4 were found empirically via grid search.

2.3 Experiments

To test our method's performance, we chose two different use cases: 1.) Finetuning MONET[6] on the expressions of the seven criteria defined in Derm7p[6,7] and 2.) finetuning MedCLIP[3] on the class and location information of lung/heart diseases in PadChest-GR.[3,25] We compared the results of SCALE to the zero-shot prediction of the original CLIP model[1] and the respective medical VLM, as well as linear probing and prompt tuning.[26] Linear probing, first introduced by Radford et al.,[1] has become a widely used baseline method.[8,26,27] In this setting, the image encoder is kept frozen, and a linear classifier is trained on top of its features, typically using the binary cross-entropy loss. The text encoder is not used during this process. For MedCLIP finetuning, we also report results for linear probing with loss weighting to account for the severe class imbalance. Furthermore, we evaluated different dataset sizes to assess the performance gains at the cost of greater labelling effort.

All models are trained for 150 epochs with a learning rate of 0.001. The batch size is adjusted adapted according to the dataset size. No augmentations were used for finetuning in the first use case. Image augmentations for finetuning are adapted from Wang et al.[3] for the finetuning of MedCLIP.

2.3.1 VLMs and Datasets

The medical foundation model MONET[6] is a CLIP-based model trained on 105,550 image-text pairs within the field of dermatology extracted from PubMed articles and medical textbooks. The image and text encoder is a ViT-L/14 and a transformer architecture with 12 self-attention layers, respectively. The input image size is 224×224 pixels, while the token vector size is limited to 77 tokens. Both encoders output a 768-dimensional embedding. MedCLIP[3] is a VLM trained on around 600,000 thoracic X-ray images from publicly available datasets. The backbone image and text encoder is BioClinicalBERT and a Swin Transformer with ImageNET pre-trained weights, respectively. Linear projection heads output feature embeddings of the size 512.

The Derm7p dataset[7] is a dataset of 1.011 patients, 252 of which suffer from melanoma. The skin lesion images are evaluated based on the 7-point skin lesion malignancy checklist and each criterion has two to eight expressions. For example, *pigmentation* is one of the evaluation criteria; Tab. 1 shows its expressions and their association with the pathology status. An increasing malignancy score determined based on the checklist correlates with the probability of the skin lesion being melanoma. The prompts for finetuning the VLMs were artificially composed by using "A photo of a skin lesion with {*concept-class-combination*}." as a template. The Derm7p dataset provides a separate test dataset.

PadChest-GR is a large-scale dataset of thoracic chest X-rays[25] with radiological reports in Spanish and English. 174 radiographic findings and 104 anatomical locations are annotated using UMLS[19] terminology. Similar to Wang et al.,[3] we evaluated the following classes: atelectasis, cardiomegaly, consolidation, and pleural effusion. Each class was assigned a coarse location description, i.e., a combination of multiple unique identifiers defined in UMLS. Locations that have less than ten cases were removed. In the following, the class and class-concept definition for finetuning MedCLIP on Padchest-GR is presented. The class-concept-combinations are followed by unique identifiers for the location definition and the number of positive cases.

```
atelectasis
├─ atelectasis left lower lobe, [(C0225758 OR C1282378) AND C0443246, C1261077], 25
├─ atelectasis retrocardiac, [Not given], 17
└─ atelectasis right lower lobe, [(C0225758 OR C1282378)] AND C0444532, C1261075], 17
cardiomegaly
└─ cardiomegaly cardiac, [C1522601], 437
consolidation
├─ consolidation left lower lobe, [(C0225758 OR C1282378) AND C0443246, C1261077], 13
└─ consolidation right lower lobe, [(C0225758 OR C1282378) AND C0444532, C1261075], 13
pleural effusion
├─ pleural effusion bilateral costophrenic angle, [C0230151 AND C0238767], 53
├─ pleural effusion bilateral pleural, [C0032225 AND C0238767], 40
├─ pleural effusion left costophrenic angle, (C0230151 AND C0443246) OR C0504100], 80
├─ pleural effusion left pleural, [C0443246 AND C0032225], 76
├─ pleural effusion right costophrenic angle(C0230151 AND C0444532) OR C0504099], 54
└─ pleural effusion right pleural, [C0444532 AND C0032225], 57
```

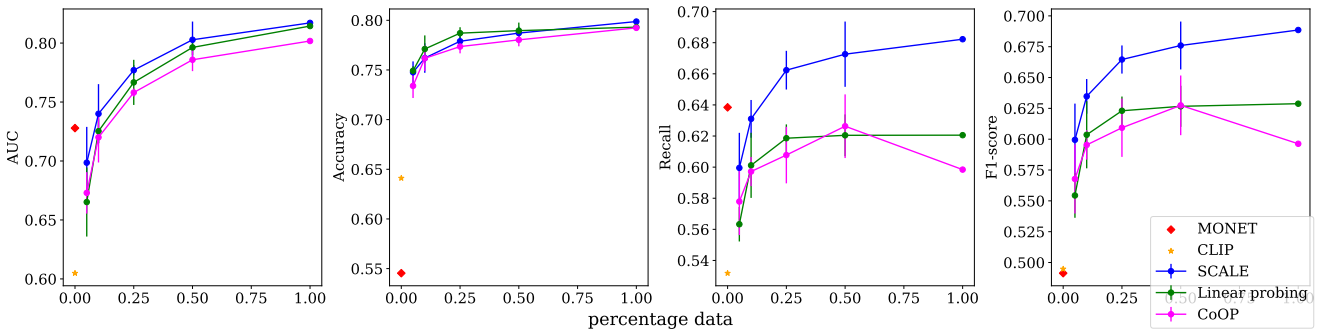
After pre-processing, the dataset consisted of 748 patients, 15% of which were allocated to the test set via a stratified split. The sentence-separated reports of Padchest-GR were used as prompts. To artificially enlarge the dataset, we mixed non-corresponding image-text pairs of the same class-concept-combination.

2.3.2 Evaluation Metrics

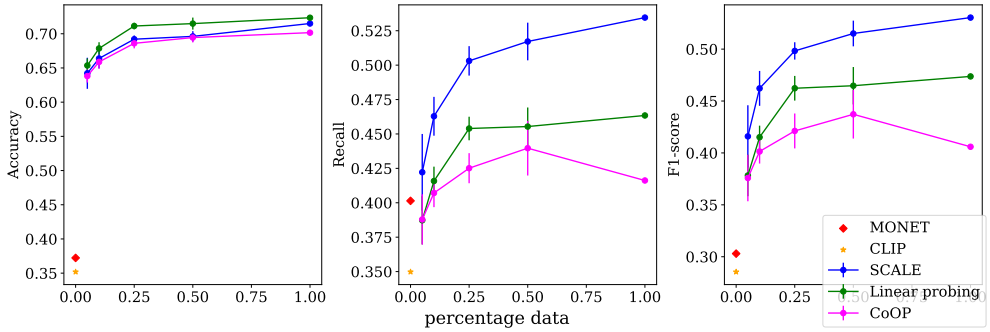
We report the mean accuracy over all classes, macro recall and macro F1-score. For binary and multi-label classification, we additionally report the AUC. For AUC calculation of the VLMs, we used the same approach as Hamamci et al.[28] : The softmax is calculated over the similarity scores given two prompt options, i.e., the class-concept-combination vs. "No finding.". The entry at the index of the ground-truth class of the resulting probability vector was used for AUC calculation. For the dermatological use case, on class-level, we report the mean over seven binary classes for each criterion for melanoma evaluation. More details can be reviewed in Tab. 1. On concept-level, the expressions of each criterion are exclusive.

3. RESULTS

As expected, the results show that the zero-shot performance of CLIP is consistently inferior to the finetuning methods starting from using 5% of training data (see Fig. 2 and Fig. 3). For the dermatological use case, the AUC and the accuracy on class-level and class-concept level is similar for all methods across different dataset sizes. However, these comparatively high values are attributable to class imbalance, and performance differences between methods become visible in the class-insensitive metrics. SCALE shows a higher recall and consequently a higher F1-score than linear probing as well as the prompt tuning method, CoOp,[26] on class- and class-concept-level. Similarly, in the results for lung and heart disease classification in Fig. 3, we identified the class imbalance in the PadChest-GR dataset to be the main cause for the poor performance of linear probing and prompt tuning in recall and F1-score. Consequently, we carried out a follow-up experiment by training the linear probing baseline with loss weighting (light green dashed line in Fig. 3). When the class imbalance is accounted for in linear probing, the method also achieves higher recall scores and similar AUC and F1-scores as the proposed method. Furthermore, we observe that the performance gain with finetuning is generally higher on class-concept- than on class-level. The t-SNE plots in Fig. 4 indicate that SCALE's text embeddings for the class–concept combinations form more coherent clusters than those of zero-shot MedCLIP.



(a) Evaluation on class-level.



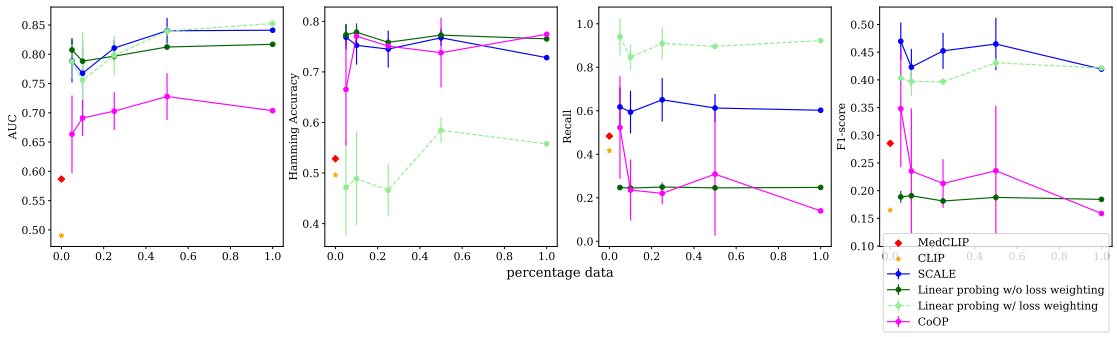
(b) Evaluation for class-concept-combinations.

Figure 2: Results for the finetuning of MONET using the Derm7p[7] dataset. The percentage of the dataset that is used for finetuning varies from 0.05, 0.1, 0.25, 0.5 to the whole dataset. When only part of the data set is used, the results are averaged over five stratified random splits. The following figures show the average results, with the error bars representing one standard deviation.

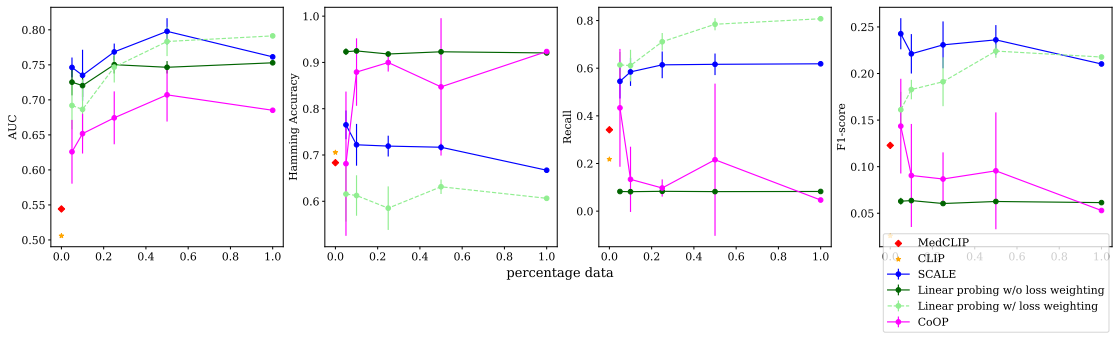
4. DISCUSSION AND CONCLUSION

In this work, we propose a finetuning approach for medical VLMs with the aim to do fine-grained class prediction by infusing CLIP adapter[8] with prior knowledge about class-concept relations. From a technical perspective, we align semantically meaningful inter-subject correspondences via supervised contrastive learning and embed the hierarchical structure of concepts and classes into the loss function. From a clinical perspective, improving the predictive performance on class-concept-level allows for zero-shot classification and image retrieval showcasing

Figure 3: Results for the finetuning of MedCLIP using the PadChest[7] dataset. The percentage of the dataset that is used for finetuning varies from 0.05, 0.1, 0.25, 0.5 to the whole dataset. When only part of the data set is used, the results are averaged over five stratified random splits. The following figures show the average results, with the error bars representing one standard deviation.

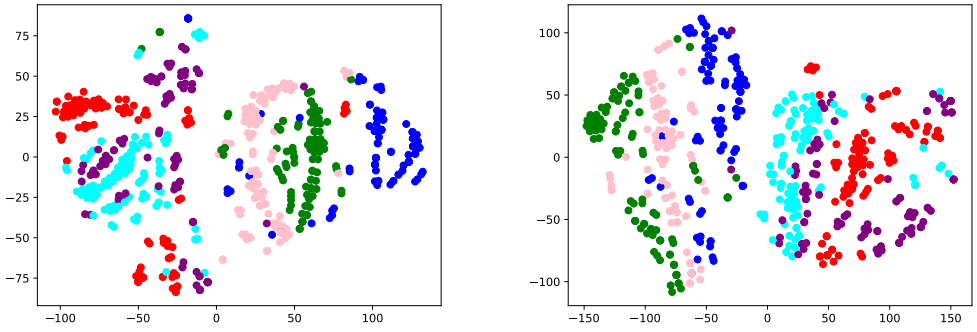


(a) Evaluation on class-level.



(b) Evaluation for class-concept-combinations.

not only the class prediction, but also the class-descriptive concepts, which give an indication of why a class is predicted. This improves the interpretability of VLMs and serves as a step towards user trust and model employment in clinical practice. Our results show that as little labelled data as 5% of the training dataset (approx. 40 data points) is needed for finetuning to strongly outperform zero-shot models on class-concept-level. Furthermore, we observe that linear probing and prompt tuning are sensitive to class imbalance, while SCALE is more robust. Despite recent advances, a substantial performance gap remains between VLMs and fully



(a) Zero-shot MedCLIP. (b) SCALE.

Figure 4: T-SNE visualization of text embeddings of 100 randomly sampled reports, that describe the class *pleural effusion* for finetuning MedCLIP on PadChest-GR.[25] Colors indicate the corresponding class-descriptive concepts.

supervised approaches for highly complex, task-specific medical imaging problems with granular class definition. As computational resources continue to grow, increasingly powerful VLMs are expected to emerge. Until such models surpass supervised training, we propose to regard VLMs themselves as a form of prior knowledge that can be leveraged for these applications. Within the methods of parameter-efficient fine-tuning of foundation models, this study has explored adapter layers and a prompt tuning technique as a comparison method. For a more comprehensive comparison, low-rank adaptation[10] could have been explored as an additional comparison method. Furthermore, in this work, we have identified the definition of the loss function for VLM finetuning as an under-explored field of research, which should be further addressed in future research. This study has shown, that leveraging prior knowledge within the loss function formulation in medical VLM finetuning on a small labelled dataset has the advantage of keeping the general knowledge of a foundation model, while simultaneously achieving competitive performance for the finetuned task.

REFERENCES

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I., "Learning Transferable Visual Models From Natural Language Supervision," *International Conference on Machine Learning* (2021).

[2] Zhao, Z., Liu, Y., Wu, H., Wang, M., Li, Y., Wang, S., Teng, L., Liu, D., Cui, Z., Wang, Q., and Shen, D., "CLIP in medical imaging: A survey," *Medical Image Analysis* **102**, 103551 (2025).

[3] Wang, Z., Wu, Z., Agarwal, D., and Sun, J., "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text," *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* **2022**, 3876–3887 (2022).

[4] Nauta, M., Hegeman, J. H., Geerdink, J., Schlötterer, J., van Keulen, M., and Seifert, C., "Interpreting and Correcting Medical Image Classification with PIP-Net," (2024).

[5] Oikarinen, T., Das, S., Nguyen, L. M., and Weng, T.-W., "Label-free concept bottleneck models," in [*International Conference on Learning Representations*], (2023).

[6] Kim, C., Gadgil, S. U., DeGrave, A. J., Omiye, J. A., Cai, Z. R., Daneshjou, R., and Lee, S.-I., "Transparent medical image AI via an image-text foundation model grounded in medical literature," *Nature Medicine* **30**(4), 1154–1165 (2024).

[7] Kawahara, J., Daneshvar, S., Argenziano, G., and Hamarneh, G., "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE Journal of Biomedical and Health Informatics* **23**(2), 538–546 (2019).

[8] Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y., "CLIP-Adapter: Better Vision-Language Models with Feature Adapters," *International Journal of Computer Vision* **132**(2), 581–595 (2024).

[9] Zhou, K., Yang, J., Loy, C. C., and Liu, Z., "Conditional Prompt Learning for Vision-Language Models," in [*2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 16795–16804, IEEE, New Orleans, LA, USA (2022).

[10] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W., "LoRA: Low-rank adaptation of large language models," in [*International Conference on Learning Representations*], (2022).

[11] Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H., "Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification," in [*Lecture Notes in Computer Science*], 493–510, Springer Nature Switzerland, Cham (2022).

[12] Huang, S.-C., Shen, L., Lungren, M. P., and Yeung, S., "GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition," in [*2021 IEEE/CVF International Conference on Computer Vision (ICCV)*], 3922–3931, IEEE, Montreal, QC, Canada (2021).

[13] Dawidowicz, G., Hirsch, E., and Tal, A., "LIMITR: Leveraging Local Information for Medical Image-Text Representation," in [*2023 IEEE/CVF International Conference on Computer Vision (ICCV)*], 21108–21116, IEEE, Paris, France (2023).

[14] Müller, P., Kaissis, G., Zou, C., and Rueckert, D., "Joint Learning of Localized Representations from Medical Images and Reports," in [*Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*], 685–701, Springer-Verlag, Berlin, Heidelberg (2022).

[15] Palepu, A. and Beam, A., "TIER: Text-Image Entropy Regularization for Medical CLIP-style models," in [*Proceedings of the 8th Machine Learning for Healthcare Conference*], 548–564, PMLR (2023).

[16] Shui, Z., Zhang, J., Cao, W., Wang, S., Guo, R., Lu, L., Yang, L., Ye, X., Liang, T., Zhang, Q., and Zhang, L., "Large-scale and fine-grained vision-language pre-training for enhanced CT image understanding," in [*International Conference on Learning Representations*], (2025).

[17] Liu, B., Lu, D., Wei, D., Wu, X., Wang, Y., Zhang, Y., and Zheng, Y., "Improving Medical Vision-Language Contrastive Pretraining With Semantics-Aware Triage," *IEEE Transactions on Medical Imaging* **42**(12), 3579–3589 (2023).

[18] Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., and Yu, L., "Multi-granularity cross-modal alignment for generalized medical visual representation learning," in [*Proceedings of the 36th International Conference on Neural Information Processing Systems*], *NIPS '22*, 33536–33549, Curran Associates Inc., Red Hook, NY, USA (2022).

[19] Bodenreider, O., "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic acids research* **32 Database issue**, D267–70 (2004).

[20] Chen, Z., Li, G., and Wan, X., "Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge," in [*Proceedings of the 30th ACM International Conference on Multimedia*], *MM '22*, 5152–5161, Association for Computing Machinery, New York, NY, USA (2022).

[21] Zhang, X., Wu, C., Zhang, Y., Xie, W., and Wang, Y., "Knowledge-enhanced visual-language pre-training on chest radiology images," *Nature Communications* **14**(1), 4542 (2023).

[22] Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W., "Medklip: Medical knowledge enhanced language-image pre-training," *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023).

[23] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D., "Supervised contrastive learning," *Advances in neural information processing systems* **33**, 18661–18673 (2020).

[24] Zhang, P. and Wu, M., "Multi-Label Supervised Contrastive Learning," *Proceedings of the AAAI Conference on Artificial Intelligence* **38**(15), 16786–16793 (2024). Number: 15.

[25] de Castro, D. C., Bustos, A., Bannur, S., Hyland, S. L., Bouzid, K., Wetscherek, M. T., Sánchez-Valverde, M. D., Jaques-Pérez, L., Pérez-Rodríguez, L., Takeda, K., Salinas-Serrano, J. M., Alvarez-Valle, J., Galant-Herrero, J., and Pertusa, A., "PadChest-GR: A Bilingual Chest X-Ray Dataset for Grounded Radiology Report Generation," *NEJM AI* **0**(0) (2025).

[26] Zhou, K., Yang, J., Loy, C. C., and Liu, Z., "Learning to Prompt for Vision-Language Models," *International Journal of Computer Vision* **130**(9), 2337–2348 (2022).

[27] Shakeri, F., Huang, Y., Silva-Rodríguez, J., Bahig, H., Tang, A., Dolz, J., and Ben Ayed, I., "Few-Shot Adaptation of Medical Vision-Language Models," in [*Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*], Linguraru, M. G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., and Schnabel, J. A., eds., 553–563, Springer Nature Switzerland, Cham (2024).

[28] Hamamci, I. E., Er, S., Almas, F., Simsek, A. G., Esirgun, S. N., Dogan, I., Dasdelen, M. F., Durugol, O. F., Wittmann, B., Amiranashvili, T., Simsar, E., Simsar, M., Erdemir, E. B., Alanbay, A., Sekuboyina, A., Lafci, B., Bluethgen, C., Ozdemir, M. K., and Menze, B., "Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography," (2024).