A Balancing Act: Optimizing Classification and Retrieval in Cross-Modal Vision Models

Judith Lefkes[1] JUDITH.LEFKES@RADBOUDUMC.NL
Clement Grisi[1] CLEMENT.GRISI@RADBOUDUMC.NL
Geert Litjens[1] GEERT.LITJENS@RADBOUDUMC.NL
[1] *Computational Pathology Group, Radboudumc, Nijmegen, Netherlands*

Editors: Under Review for MIDL 2025

Abstract

Despite the promising capabilities of vision-language models (VLMs) across diverse tasks, recent studies reveal that they struggle with the fundamental task of image classification. In this study, we explore leveraging state-of-the-art task-specific classification models as a foundation for VLMs, aiming to preserve strong classification performance. Specifically, we assess the impact of contrastive tuning to enable cross-modal retrieval capabilities on a Hierarchical Image Pyramid Transformer (HIPT) trained for prostate cancer grading in Whole-Slide Images (WSIs) and a ViT-Base model trained for multi-label classification on natural images. Our results demonstrate that contrastive fine-tuning creates a clear trade-off: classification accuracy rapidly deteriorates toward zero as vision-text alignment improves. By balancing the two objectives in the loss function during fine-tuning, we achieve competitive slide-level retrieval performance while maintaining classification accuracy.

Keywords: Multi-task Learning, Vision-Language Models, Representation disentanglement, Computational Pathology

1. Introduction

The field of computational pathology is seeing an increase in the development of foundation models (FMs) (Vorontsov et al., 2024; Ikezogwo et al., 2023). Large-scale pretraining using self-supervised learning (SSL) on thousands of histopathological slides spanning diverse tissue types and diseases can provide foundation models with advantages over task-specific models. They can serve as a general foundation for various downstream tasks in pathology, such as cancer subtyping and prognostication (Wang et al., 2024; Chen et al., 2024). Vision-language models (VLMs), a subset of foundation models incorporating textual data from sources like pathology reports, educational materials, or PubMed, can learn cross-modal associations (Lu et al., 2023). These models have demonstrated promising capabilities in tasks such as cross-modal retrieval (Lu et al., 2023), image captioning (Lu et al., 2023; Shaikovski et al., 2024), and report generation (Tran et al., 2024).

Despite their successes, recent computer vision research highlights VLMs' critical limitations. In particular, VLMs significantly underperform on standard image classification benchmarks compared to state-of-the-art (SOTA) task-specific classification models (Laurençon et al., 2024; Karamcheti et al., 2024; Zhang et al., 2024; Tong et al., 2024; Zhai et al., 2023). Zhang et al. (2024) attribute this shortfall primarily to the limited availability of classification-focused data during pretraining of VLMs. Zhai et al. (2023) demonstrate

that fine-tuning VLMs with classification-focused data enhances in-domain performance but causes catastrophic forgetting, leading to reduced performance on out-of-domain datasets and compromised generalizability.

In high-stakes domains like medicine, where diagnosis guides treatment decisions and directly impacts patient outcomes, even slight declines in classification performance can have serious consequences. This raises a key question: can task-specific vision models be adapted for multi-modal tasks without compromising their classification performance? How much classification-specific information do we sacrifice in favor of cross-modal alignment?

To address this question, we begin with SOTA task-specific image classification models and explore the impact of contrastive tuning for enabling cross-modal tasks like image-to-text retrieval. Without any mitigation strategy, we hypothesize that the model will suffer from catastrophic forgetting while adapting to the cross-modal task. To mitigate forgetting, we introduce a balancing parameter, λ, which modulates the relative emphasis on classification and vision-language alignment in the loss function.

We summarize our contributions as follows:

1. We show that contrastive fine-tuning without a classification objective leads to catastrophic forgetting, where classification accuracy deteriorates rapidly in favor of vision-text alignment in general vision and the medical domain.

2. To address this trade-off, we propose fine-tuning with a dual-objective loss function weighted by a balancing parameter, λ, which controls the trade-off between classification and contrastive objectives.

3. We show that λ selection is task-specific and that we can achieve competitive retrieval performance through careful tuning while preserving classification accuracy on a prostate cancer grading task.

2. Methods

2.1. Experimental setup

To demonstrate that our results are applicable and transferable to both natural and medical images, we conduct experiments on two distinct datasets: the Microsoft Common Objects in Context (COCO) (Lin et al., 2015) dataset and a curated medical dataset of prostate biopsies and corresponding pathology reports.

We start with a high-performing vision model for a specific classification task and a frozen language encoder to test our hypothesis that classification performance is traded away when fine-tuning for cross-modal performance. We then use a dual-objective loss function that weights a classification and contrastive objective by a parameter λ, defined as follows:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{contrastive}} + (1 - \lambda)\mathcal{L}_{\text{classification}}$$

Thus, λ of 0.0 implies disregarding the contrastive objective and continuing fine-tuning for classification, while a 1.0 is equivalent to purely focusing on the contrastive objective. We hypothesize that the higher the λ, the more classification performance you lose. Additionally, we assume that the best value for λ is task or dataset-specific.

We analyze the trade-off between classification and cross-modal alignment by tracking validation performance metrics per epoch for different lambdas during contrastive tuning. Moreover, we optimized for several epochs over early stopping to assess the trade-offs rather than fully maximize peak performance. Our code is available on github.

2.2. COCO Experiments

Dataset
We select 30,000 image-caption pairs from the 2014 MS COCO release for contrastive tuning. Of these, 4,952 pairs are held out for independent testing, while the remaining pairs are divided into five cross-validation folds. Figure 1A shows an example of an image-caption pair.

Models
The COCO experiments use a Vision Transformer (ViT)-base architecture, *google/vit-base-patch16-224* (Wu et al., 2020) fine-tuned for multi-label classification on the 80 classes in the dataset. Fine-tuning details are provided in Appendix A.

We experimented with two publicly available SentenceTransformer models (Reimers and Gurevych, 2019) for the language encoder. We report results in the main paper using the RoBERTa base model (*roberta-base-nli-stsb-mean-tokens*) and present additional results using the MPNet model architecture (*multi-qa-mpnet-base-dot-v1*) in the Appendix C.0.1.

Evaluation Metrics
We evaluate multi-label classification using mean average precision (mAP) and vision-language alignment through image-to-text retrieval. Retrieval performance is measured by *Recall@K*, where the top K captions are retrieved from 4,690 validation captions based on the cosine similarity between image and text embeddings. A retrieval is considered correct if at least one of the K-retrieved captions matches any of the five reference captions associated with the image.

Implementation Details
Experiments are run over 50 epochs using the Binary Cross-Entropy (BCE) loss for the multi-label classification task and the CLIP Loss (Radford et al., 2021) for vision-language alignment. Optimization is performed with the AdamW optimizer, employing a learning rate of 1×10^{-4}, a weight-decay parameter of 0.001, and a batch size of 64.

2.3. Prostate biopsy grading experiments

Dataset
We curated a dataset of 425 WSIs containing a single prostate biopsy, their ISUP grade, and a report from the Radboud University Medical Center, Nijmegen, the Netherlands. Each report consists of a microscopy and conclusion section. An example is shown in Figure 1B, and the label distribution of the dataset in Figure 1C. We reserve a test set of 35 cases and partition the remaining data into five stratified cross-validation folds based on ISUP grade.

Models

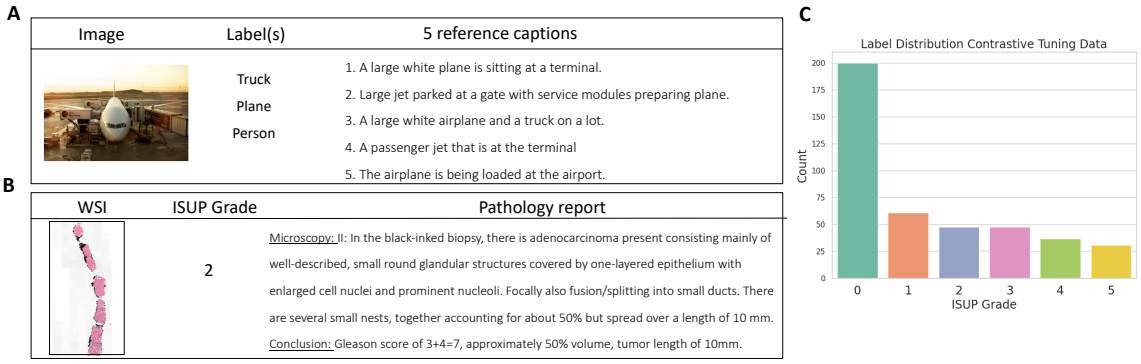

Figure 1: **(A)** An example case from the COCO dataset comprising a natural image, one or multiple label(s), and five reference captions. **(B)** An example case from the prostate biopsy data comprising a thumbnail of a WSI, its corresponding ISUP grade, and the pathology report. **(C)** Label distribution of the prostate biopsy data.

For the task-specific vision model, we trained a hierarchical vision transformer (HIPT), pretrained with the DINO framework on the PANDA dataset of $11,554$ $H\&E$-stained prostate WSIs (Bulten et al., 2022). This model achieves state-of-the-art performance in multi-class ISUP grade classification with a quadratic kappa score of 0.892 on the PANDA test set (Grisi et al., 2023). Given the small tuning dataset size, we freeze the two patch and region transformers of our HIPT model and update only the weights of the final transformer.

For the language encoder, we report results in the main paper using a model pretrained on Dutch clinical reports and fine-tuned for the task of predicting ISUP grade from the microscopic sections of a pathology report (see Appendix B for details) (Bosma, 2024). Additionally, results using the *BioBERT* model (Lee et al., 2020), pretrained on English biomedical text are presented in the Appendix C.0.2 for comparison.

Evaluation Metrics

We evaluate ISUP grade classification performance on prostate biopsies using the quadratic kappa score (κ^2). For retrieval, we introduce a new metric, *retrieval* κ^2, to measure WSI-level image-to-text retrieval. *Retrieval* κ^2 assesses the agreement between each slide's original labels and the labels of the top-one retrieved report by calculating Cohen's quadratic kappa.

Implementation Details

For ISUP grade classification, we frame the task as a regression problem to better capture the ordinal nature of ISUP scores (Grisi et al., 2023) and use the Mean Squared Error (MSE) loss for classification and the Triplet loss for cross-modal alignment. We run experiments for a maximum of 30 epochs, using the AdamW optimizer with a learning rate of 1×10^{-5}, a weight decay of 0.001, and a StepLR scheduler with a step size of five and a decay factor of 0.5. We use a batch size of 1 with gradient accumulation over 16 steps.

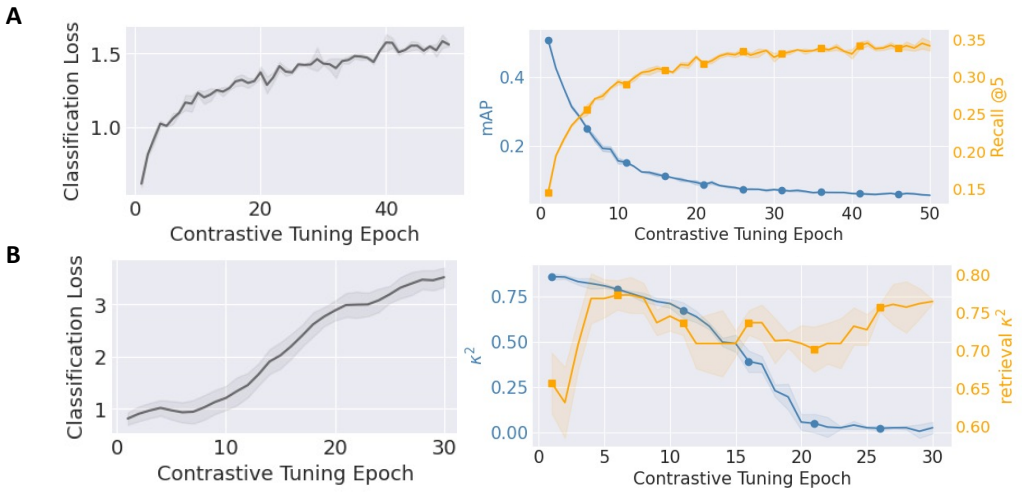

Figure 2: Validation performance metrics during contrastive tuning using $\lambda = 1.0$ for the COCO dataset in **(A)** and the prostate biopsy grading experiments in **(B)**. Lines represent the medians across five folds, with shaded areas indicating the interquartile range.

3. Results

3.1. COCO

Baseline
Before contrastive tuning, our vision encoder achieves an mAP of 0.77 on the independent test set, a median classification loss of 0.049, and a median mAP of 0.768 on the validation sets.

Contrastive tuning without classification objective ($\lambda = 1.0$)
We evaluate the most natural choice for the hyperparameter λ, specifically $\lambda = 1.0$ in Figure 2A. We observe a clear trade-off: classification performance declines immediately after the first epoch, as reflected by a steep increase in classification loss. At the same time, alignment improves significantly, as indicated by a moderate *Recall@K* achieved at around 25 epochs of fine-tuning. Finally, the mAP reaches zero after approximately 30 epochs, reflecting the complete loss in classification capabilities when tuning without a classification objective.

Balancing classification and alignment
To address the trade-off, we evaluate intermediate values of the hyperparameter λ to balance classification and contrastive objectives during tuning. Figure 3A shows the results. The training classification loss decreases linearly across all values of λ with lower values of λ (e.g., $\lambda = 0.1$), achieving lower final loss as they prioritize the classification compared to higher λ values (e.g., $\lambda = 0.9$) which favor vision-language alignment. In contrast, the validation classification loss increases more rapidly for lower λ values, suggesting that the

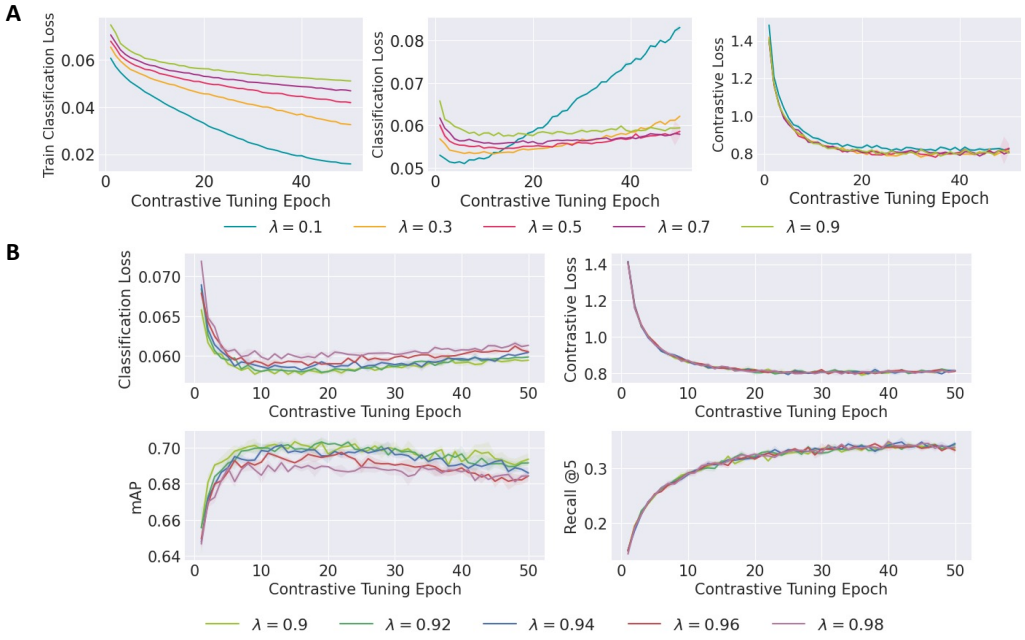

Figure 3: Impact of λ on the classification-alignment trade-off for COCO with **(A)** $\lambda \in [0.1, 0.9]$ and **(B)** $\lambda \in [0.9, 1.0]$.

model starts overfitting on the classification task. The validation contrastive loss converges quickly and displays similar trajectories across all λ values, highlighting that a stronger emphasis on classification does not severely hinder contrastive learning performance.

Optimizing λ to minimize catastrophic forgetting
In our third experiment, we redefine λ as the range [0.9, 1.0) to isolate its impact from the previously observed overtraining effect, as shown in Figure 3B. Selecting λ closer to 1.0 should maximize multi-modal alignment, mitigating overfitting and identifying the point at which classification performance begins to decline. Indeed, classification loss increases, but less sharply than when no mitigation is applied ($\lambda = 1.0$), and this is accompanied by a slight decline in mAP. Contrastive loss and retrieval performance remain largely unaffected by the choice of λ, stabilizing around 25 epochs. Importantly, lower values (e.g., 0.9) achieve marginally better mAP compared to higher values like 0.98, indicating values around, e.g., $\lambda = 0.9$ may be ideal for this task as they maintain the highest classification performance while obtaining similar retrieval performance.

3.2. Prostate biopsy grading

Baseline The pretrained HIPT model achieves a $\kappa^2 = 0.8$ on the independent test set and a median $\kappa^2 = 0.839$ with a median classification loss of 0.857 across five validation folds.

Contrastive tuning without classification objective ($\lambda = 1.0$)

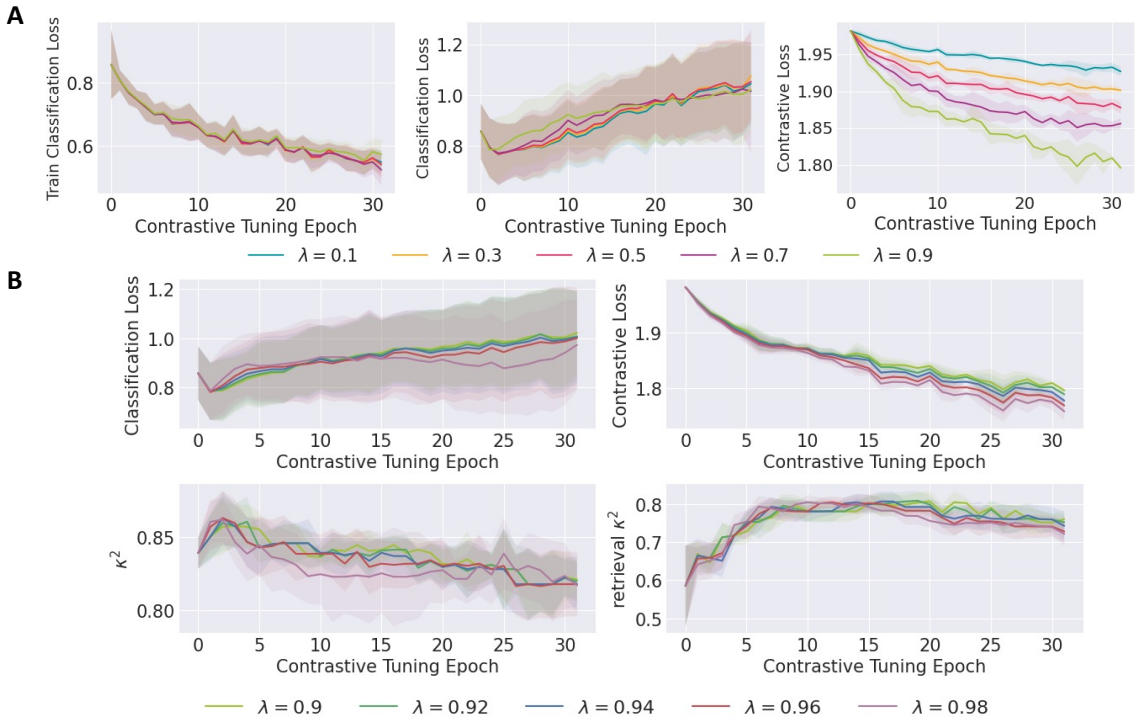

Figure 4: Impact of λ on the classification-alignment trade-off for the prostate biopsy grading experiments with **(A)** $\lambda \in [0.1, 0.9]$ and **(B)** $\lambda \in [0.9, 1.0]$.

As illustrated in Figure 2B, we observe a rise in classification loss alongside continued contrastive alignment optimization, confirming a similar trade-off for prostate cancer grading as in COCO. Consistent with prior observations, fine-tuning with $\lambda = 1.0$ results in a complete loss of classification performance for prostate biopsies within 20 epochs, trading it for a retrieval κ^2 of approximately 0.8.

Balancing classification and alignment Using $\lambda \in [0.1, 0.9]$ for the medical task results in a consistent rise in classification loss with minimal variation across λ values as illustrated in Figure 4A. This trend, observed after just a few epochs, indicates a potential overtraining effect consistent across all λ values. However, the loss stabilizes around 1.0, significantly lower than the approximately 3.5 observed with $\lambda = 1.0$ after 30 epochs. Both training and validation classification losses exhibit higher variability compared to natural images. In contrast, the contrastive loss shows noticeable differences, as lower λ values result in significantly lower final contrastive loss, suggesting that higher values may be more favorable for maximizing alignment.

Optimizing λ to minimize catastrophic forgetting
Figure 4B displays results using $\lambda \in [0.9, 1.0]$. Regarding the losses for the two objectives, there is no clear difference between the intermediate and higher ranges of λ. Lower λ values,

such as 0.9, appear more advantageous, achieving comparable retrieval performance while maintaining higher classification accuracy. However, the high variability across folds complicates the precise interpretation of performance scores. Still, classification performance declines are mitigated for all $\lambda \neq 1.0$, maintaining a competitive κ score of 0.82 after 30 epochs.

4. Discussion

In the medical domain, where accurate classification underpins critical tasks such as clinical decision-making and treatment planning, task-specific algorithms remain the standard for AI systems implemented in the clinic. This paper explored whether task-specific classification models can serve as a foundation for multi-modal systems, aligning cross-modal objectives without sacrificing classification performance.

Our findings indicate that contrastive tuning of a task-specific vision model without a classification objective results in catastrophic forgetting. The classification performance declined to nearly zero within fewer than 30 epochs in both the COCO experiments and the medical task as the model increasingly prioritized vision-text alignment. These findings align with the catastrophic forgetting literature, where neural networks lose previously learned information when adapting to new objectives. They may also explain why VLMs often fail to surpass SOTA vision classifiers in classification tasks.

We proposed a simple yet effective approach to address this trade-off by integrating a classification objective into the loss function during contrastive tuning. By carefully tuning the weighting factor λ, we effectively reduced the decline in classification performance from a complete 100% drop ($\lambda = 1.0$) to just 10% with $\lambda = 0.9$ after 50 epochs, thus retaining approximately 90% of the baseline mAP performance in COCO. Similarly, contrastive tuning using $\lambda = 0.9$ for the prostate cancer grading task reduced the total κ^2 drop to approximately 3% while achieving acceptable retrieval performance within just a few epochs. The consistent trend is independent of the language encoder, as shown in Appendices C.0.1 and C.0.2. Moreover, our experiments demonstrate that the optimal value of λ varies depending on the task.

Our study has some limitations. First, a single dataset per domain was used. Second, although we propose a simple solution to mitigate the loss of classification performance, a more thorough investigation and the development of more sophisticated methods could improve and simplify the management of the trade-off between classification and retrieval.

Third, while the variability in the prostate cancer grading dataset is relatively high, we anticipate that this variability could be reduced and that higher overall retrieval performance will be achieved with a larger dataset.

In summary, this study calls for a renewed focus on catastrophic forgetting as a critical challenge in multi-task model development in the medical field. By developing strategies that ensure that fundamental classification capabilities are preserved, we can pave the way for building more robust models that are better suited for clinical implementation.

References

Joeran Bosma. dragon-roberta-base-domain-specific. 2024. doi: 10.57967/HF/2169. URL https://huggingface.co/joeranbosma/dragon-roberta-base-domain-specific. Publisher: Hugging Face Version Number: 7f8facc.

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester van Boven, Robert Vink, Christina Hulsbergen-van de Kaa, Jeroen van der Laak, Mahul B. Amin, Andrew J. Evans, Theodorus van der Kwast, Robert Allan, Peter A. Humphrey, Henrik Grönberg, Hemamali Samaratunga, Brett Delahunt, Toyonori Tsuzuki, Tomi Häkkinen, Lars Egevad, Maggie Demkin, Sohier Dane, Fraser Tan, Masi Valkonen, Greg S. Corrado, Lily Peng, Craig H. Mermel, Pekka Ruusuvuori, Geert Litjens, and Martin Eklund. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28(1):154–163, January 2022. ISSN 1546-170X. doi: 10.1038/s41591-021-01620-2. URL https://www.nature.com/articles/s41591-021-01620-2. Publisher: Nature Publishing Group.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3. URL https://www.nature.com/articles/s41591-024-02857-3. Publisher: Nature Publishing Group.

Clément Grisi, Geert Litjens, and Jeroen van der Laak. Hierarchical Vision Transformers for Context-Aware Prostate Cancer Grading in Whole Slide Images, December 2023. URL http://arxiv.org/abs/2312.12619. arXiv:2312.12619 [cs].

Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1M: One Million Image-Text Pairs for Histopathology, October 2023. URL http://arxiv.org/abs/2306.11207. arXiv:2306.11207 [cs].

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models, May 2024. URL http://arxiv.org/abs/2402.07865. arXiv:2402.07865 [cs].

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, May 2024. URL http://arxiv.org/abs/2405.02246. arXiv:2405.02246 [cs].

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, February 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, February 2015. URL http://arxiv.org/abs/1405.0312. arXiv:1405.0312.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, Georg Gerber, Anil V. Parwani, and Faisal Mahmood. Towards a Visual-Language Foundation Model for Computational Pathology, July 2023. URL http://arxiv.org/abs/2307.12914. arXiv:2307.12914 [cs].

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, August 2019. URL http://arxiv.org/abs/1908.10084. arXiv:1908.10084 [cs].

George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D. Kunz, Juan A. Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, Matthew Hanna, Michal Zelechowski, Julian Viret, Neil Tenenholtz, James Hall, Nicolo Fusi, Razik Yousfi, Peter Hamilton, William A. Moye, Eugene Vorontsov, Siqi Liu, and Thomas J. Fuchs. PRISM: A Multi-Modal Generative Foundation Model for Slide-Level Histopathology, May 2024. URL http://arxiv.org/abs/2405.10254. arXiv:2405.10254 [eess].

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs, April 2024. URL http://arxiv.org/abs/2401.06209. arXiv:2401.06209 [cs].

Manuel Tran, Paul Schmidle, Sophia J. Wagner, Valentin Koch, Valerio Lupperger, Annette Feuchtinger, Alexander Böhner, Robert Kaczmarczyk, Tilo Biedermann, Kilian Eyerich, Stephan A. Braun, Tingying Peng, and Carsten Marr. Generating highly accurate pathology reports from gigapixel whole slide images with HistoGPT, March 2024. URL http://medrxiv.org/lookup/doi/10.1101/2024.03.15.24304211.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Juan Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David Klimstra, Brandon Rothrock, and Thomas J. Fuchs. Virchow: A Million-Slide Digital Pathology Foundation Model, January 2024. URL http://arxiv.org/abs/2309.07778. arXiv:2309.07778 [eess].

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, Fang Wang, Yulong Peng, Junyou Zhu, Jing Zhang, Christopher R. Jackson, Jun Zhang, Deborah Dillon, Nancy U. Lin, Lynette Sholl, Thomas Denize, David Meredith, Keith L. Ligon, Sabina Signoretti, Shuji Ogino, Jeffrey A. Golden, MacLean P. Nasrallah, Xiao Han, Sen Yang, and Kun-Hsing Yu. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, October 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07894-z. URL https://www.nature.com/articles/s41586-024-07894-z. Publisher: Nature Publishing Group.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual Transformers: Token-based Image Representation and Processing for Computer Vision, November 2020. URL http://arxiv.org/abs/2006.03677. arXiv:2006.03677 [cs].

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the Catastrophic Forgetting in Multimodal Large Language Models, December 2023. URL http://arxiv.org/abs/2309.10313. arXiv:2309.10313 [cs].

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are Visually-Grounded Language Models Bad at Image Classification?, November 2024. URL http://arxiv.org/abs/2405.18415. arXiv:2405.18415 [cs].

Appendix A. Fine-tuning Details for the Task-specific Vision Encoder in COCO

We utilize the MS COCO 2014 dataset, which consists of 123,287 images, each paired with five reference captions (training + validation). For vision-only fine-tuning, we randomly select 93,813 image-label pairs stratified across 80 classes, transforming the *google/vit-base-patch16-224* architecture into a task-specific multi-label classification model. The remaining 29,474 image-caption pairs are reserved for the contrastive tuning experiments in the main paper. The data is split into training, validation, and test sets (80/10/10). Fine-tuning is performed for a maximum of 50 epochs using the binary cross-entropy (BCE), with early stopping applied (patience = 10). Optimization is conducted using the the AdamW optimizer with a learning rate of $1e-4$, a weight decay of 0.001, and a batch size of 64. The fine-tuned model achieves a mAP of 0.77 on the test set, and the resulting weights are used as initialization for contrastive tuning.

Appendix B. Fine-Tuning Details for the Language Encoder Pretrained on Dutch Medical Reports

We further fine-tuned the *joeranbosma/dragon-bert-base-domain-specific* (Bosma, 2024) model for the task of predicting ISUP grade from the microscopic sections of pathology reports. This fine-tuning ensures that the [CLS] token generates meaningful sentence embeddings of dimension 768, as it is not inherently optimized for this during MLM pretraining. Additionally, fine-tuning was performed to meet the requirements of contrastive learning, where the output dimensions of the vision and language encoders need to be aligned.

Appendix C. Impact of Language Encoder Choice: Results with BioBERT and MPNet

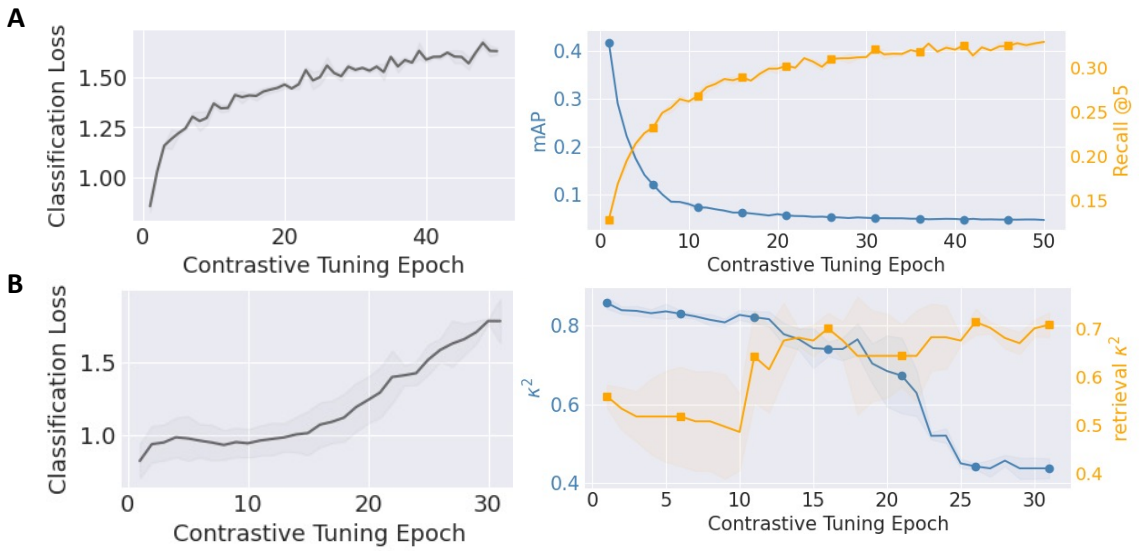

Figure 5: Validation performance metrics during contrastive tuning with $\lambda = 1.0$. Panel **(A)** presents classification loss, mAP, and Recall@5 for the COCO dataset, where text embeddings are computed using the *MPNet* model. Panel **(B)** shows validation classification loss, retrieval κ^2, and classification κ^2 for prostate cancer grading experiments, where report embeddings are derived from the *BioBERT* model. Both panels illustrate a clear trade-off, where classification performance is sacrificed in exchange for improved retrieval. In all figures, metrics are reported starting from the first epoch of fine-tuning. Lines represent the median across five folds, with shaded areas indicating the interquartile range (IQR).

C.0.1. Contrastive Tuning Experiments on COCO using **MPNet** as a Language Encoder

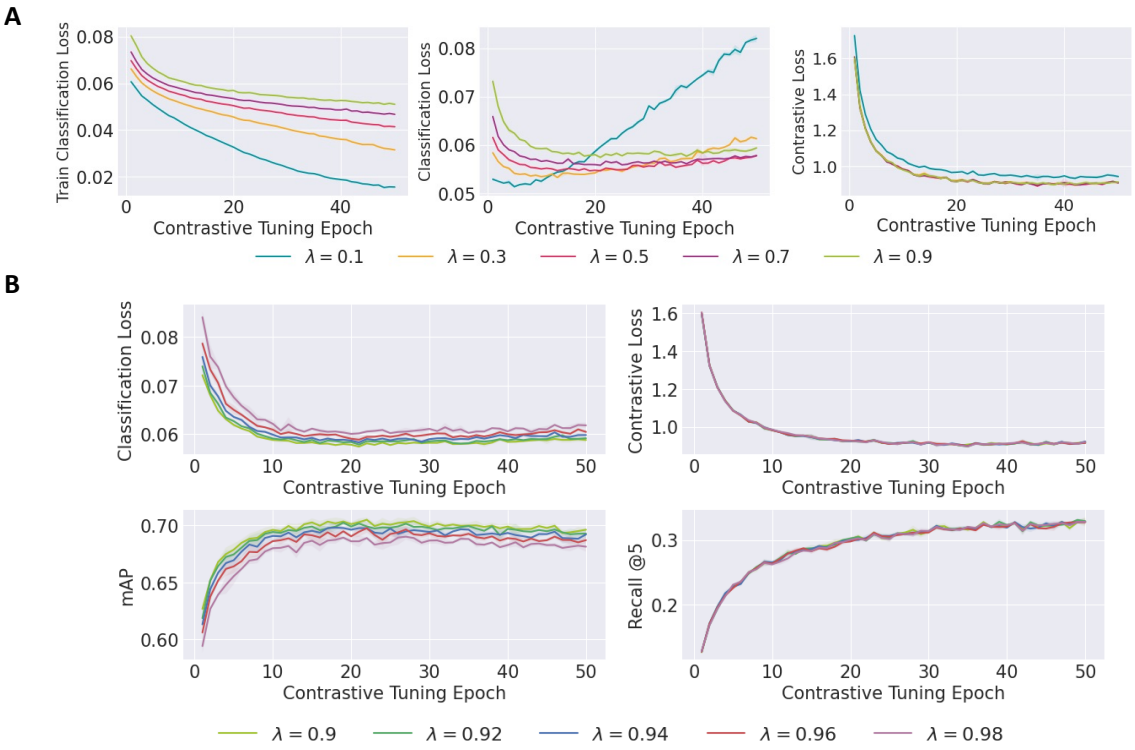

Figure 6: Impact of λ on the classification-alignment trade-off for COCO with **(A)** $\lambda \in [0.1, 0.9]$ and **(B)** $\lambda \in [0.9, 1.0]$. We used the same Vit-Base model for the vision encoder while generating text embeddings using the frozen *multi-qa-mpnet-base-dot-v1* model as the language encoder.

C.0.2. Contrastive Tuning Experiments on the Prostate Biopsy Data using **BioBERT** as a Language Encoder

For the experiments utilizing *BioBERT*, the original Dutch reports were translated into English using the *Nous-Hermes-2-Mistral-7B-DPO.Q4-0.gguf* model and the GPT4ALL python library as *BioBERT* is primarily trained on English biomedical text.

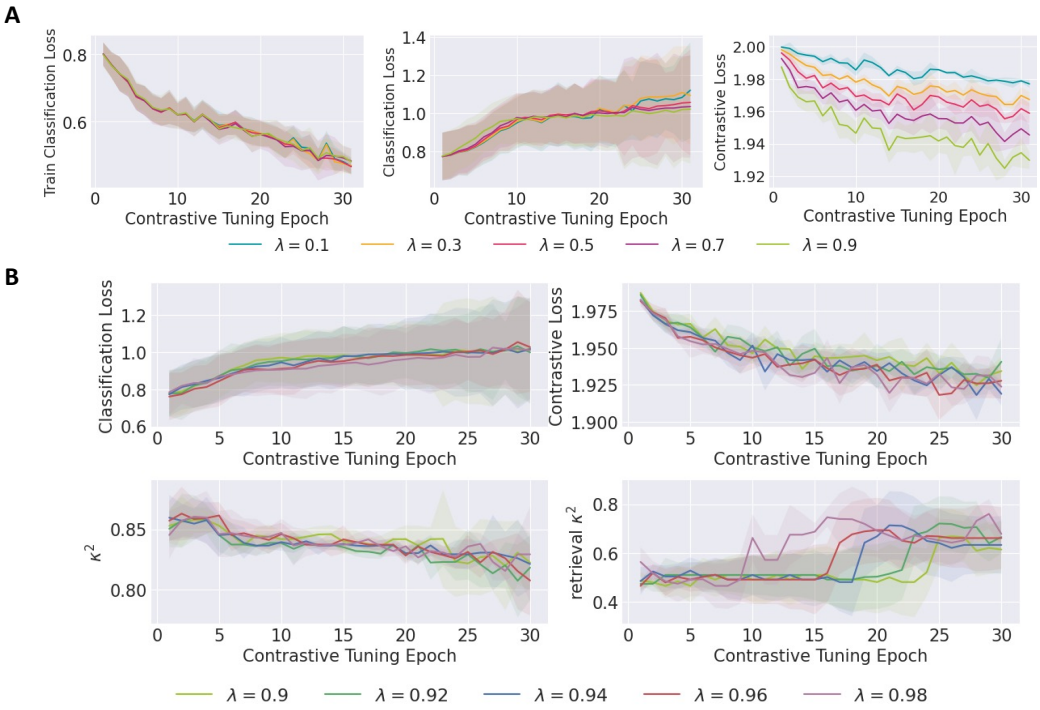

Figure 7: Impact of λ on the classification-alignment trade-off for the prostate cancer grading task with **(A)** $\lambda \in [0.1, 0.9]$ and **(B)** $\lambda \in [0.9, 1.0]$. We use the same HIPT model as the vision encoder while generating text embeddings with the frozen *BioBERT* model as the language encoder.