# Benchmarking Parameter Efficient Adaptation of Vision Language Models on Pathology

**Shivam Rajendra Rai Sharma**[1]* **Xiaoguang Zhu**[2]* **Luca Cerny Oliveira**[2]
**Kartik Patwari**[2] **La Rissa Vasquez**[3] **David Garcia**[3] **Louise Nicole C. Sevilla**[3]
**Brittany N. Dugger**[3] **Chen-Nee Chuah**[2]

[1]Department of Computer Science, University of California, Davis,
[2]Department of Electrical and Computer Engineering, University of California,
Davis, Davis, CA 95616
[3]Department of Pathology and Laboratory Medicine, University of California,
Davis, Sacramento, CA 95817

{srsrai,xgzhu, lcernyo, kpatwari, chuah}@ucdavis.edu,
{livasquez, drdgarcia, lcsevilla, bndugger}@health.ucdavis.edu

## Abstract

Generalist vision–language models (VLMs) struggle on histopathology tasks due to domain gaps and scarce labels. Pathology VLMs (PFMs) also fall short despite costly pretraining. Parameter-efficient fine-tuning (PEFT) offers a scalable lightweight approach to quickly adapt large pretrained models to target histopathology tasks. We present the first benchmark of PEFT methods when applied to VLMs/PFMs for histopathology tasks. We categorize existing PEFT methods based on adaptation modality, strategy and locus. We curate a novel neuropathology dataset for detecting neurofibrillary tangles (NFTs), a hallmark of Alzheimer's Disease, capturing annotator variability to evaluate reliability and alignment. Experiments across prostate cancer, colorectal cancer, and neuropathology tasks show that with full data, PEFT-adapted generalist VLMs rival adapted PFMs, but fall short in few shot settings due to label scarcity, terminology mismatch, and modality-specific biases. Visualization further reveals that models such as CONCH+MMRL focus on NFT within annotated boxes, improving interpretability in single-NFT cases, but their performance diminishes in complex multi-NFT scenarios. Together, our benchmark and dataset highlight PEFT as a scalable strategy, but also indicate the need for richer interpretability metrics and improved multimodal reasoning to handle complex cases.

## 1 Introduction

Large vision–language models (VLMs) such as CLIP [Radford et al., 2021] have shown remarkable zero-shot generalization across natural image domains, inspiring interest in their application to medical image analysis. In histopathology, pathologists rely on careful interpretation of complex tissue patterns to diagnose and grade disease—tasks that require fine-grained visual discrimination under limited annotation regimes. For example, accurate detection of neurofibrillary tangles (NFTs), in human brain, is essential for Alzheimer's research [Alafuzoff et al., 2008, Dugger and Dickson, 2017], yet manual annotation of whole-slide images (WSIs) remains time-consuming and resource-intensive [Ghandian et al., 2024]. Compared to natural images, WSIs present unique challenges:

*These authors contributed equally to this work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

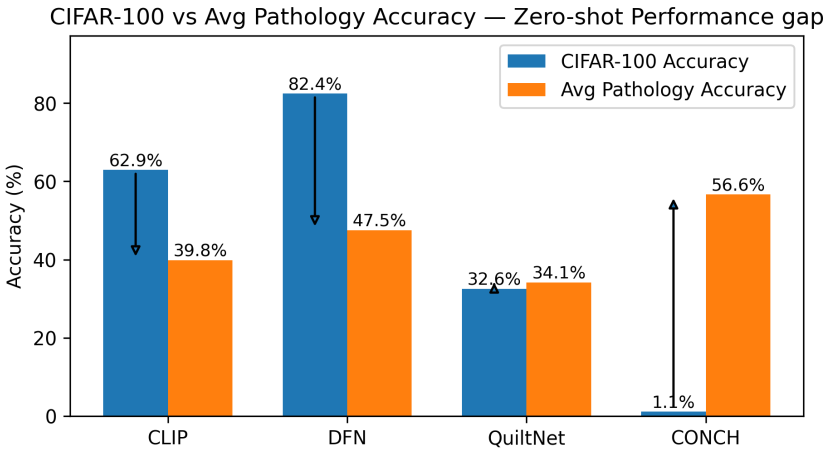

Figure 1: Zero-shot performance comparison of vision-language models (VLMs) across natural and pathology domains. Bars show average accuracy of natural image classification task using CIFAR-100 and across three representative histopathology tasks. Downward arrows indicate performance drops when transferring from natural to pathology domains, while upward arrows indicate models that perform better on histopathology tasks.

gigapixel resolution, texture-dominated morphology, and highly imbalanced disease distributions, all of which create a substantial domain gap that hinders direct transfer of general-purpose VLMs [Lai et al., 2024]. As shown in Fig. 1 and Table A1, zero-shot VLMs like CLIP and DFN [Fang et al., 2023] deteriorate in classification accuracy averaged across three pathology tasks - Gleason grading in prostate cancer, NFT detection in Alzheimer's brain tissue, and colorectal cancer tissue classification (described in Section 2). To mitigate this gap, pathology-specific VLMs such as PLIP [Huang et al., 2023], QuiltNet [Ikezogwo et al., 2023], BioMedCLIP [Zhang et al., 2023], and CONCH [Lu et al., 2024] have been introduced to improve performance through pretraining with pathology datasets. However, their heavy compute and data demands hinder clinical deployment— training CLIP required 12 days on 256 V100 GPUs for 400M samples, while QuiltNet used 4 A40 GPUs for 40 epochs on 1M pairs, yet neither generalized to histopathology tasks, underscoring the high cost of domain-specific pretraining. This motivates parameter-efficient fine-tuning (PEFT), which aligns frozen VLMs to the target task through lightweight modules that update only a small fraction of parameters, hence more efficient.

**PEFT Prior Work:** A variety of PEFT approaches have been proposed and shown to achieve strong efficiency–accuracy trade-offs on natural image benchmarks. Linear probing [Radford et al., 2021], a simple and efficient baseline — trains only a classifier on frozen CLIP features. CoOp [Zhou et al., 2022b] learns continuous contextual text tokens beyond class names; CoCoOp [Zhou et al., 2022a] conditions those tokens on image features to aid generalization. ProVP [Xu et al., 2025] extends prompting to vision by inserting learnable visual tokens across transformer layers, regularized to align with frozen features. Adapter-based strategies like CLIP-RFC (CLIPath) [Lai et al., 2023] introduce residual connections on top of vision embeddings to stabilize adaptation. Multi-modal approaches include MaPLe[Khattak et al., 2023], which jointly learns coupled prompts for shallow layers of both text and vision encoders to improve alignment, while MMA [Yang et al., 2024], inserts adapters into deeper layers of both modalities for finetuning. Finally, representation-regularized approaches such as MMRL [Guo and Gu, 2025a] and its streamlined variant MMRL++ [Guo and Gu, 2025b] optimize adapted features with distributional and consistency constraints to maintain robustness and stability under limited data. However, systematic evaluations of PEFT for pathology tasks are lacking [Mai et al., 2025], while other benchmark studies primarily focus on pathology foundation models pretrained with large datasets and overlook PEFTs [Lee et al., 2025, Xiong et al., 2025, Bareja et al., 2025]. This leaves an open question: *can PEFT enable reliable and interpretable adaptation of VLMs to pathology tasks without relying on expensive domain-specific pretraining?*

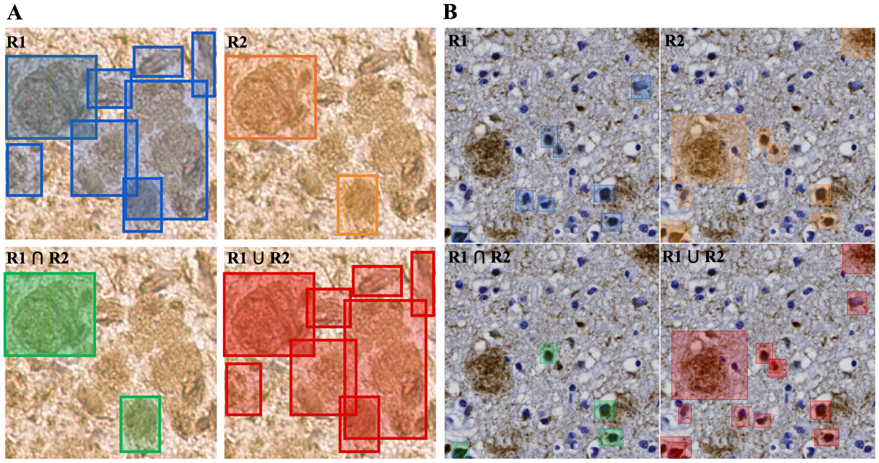

Figure 2: Inter-annotator agreement on tau pathology across stains and scanners. **(A)** PHF-1 stained brain section scanned on an Aperio AT2 (0.23 µm/pixel). **(B)** AT8–stained brain section scanned on a Zeiss Axio Scan.Z1 (0.11 µm/pixel). **Top row (A, B):** annotations from two independent human raters — R1 shown with blue bounding boxes and R2 with orange bounding boxes. **Bottom row:** consensus annotations — $R_1 \cap R_2$ (agreement; green) and $R_1 \cup R_2$ (inclusive set; red). Intersection regions serve as high-confidence references, while union regions provide comprehensive targets for evaluating model focus on neurofibrillary tangles and reliability in task alignment.

Table 1: Head-to-head winners across four VLMs (best method per VLM chosen by full-data AUC).

| **Task** | **CLIP Best (Method, AUC)** | **DFN Best (Method, AUC)** | **QuiltNet Best (Method, AUC)** | **CONCH Best (Method, AUC)** | **Winner (ΔAUC)** |
|---|---|---|---|---|---|
| SICAPv2 (Gleason) | MMRL (92.73) | MaPLe (93.14) | MaPLe (94.14) | CLIPath (94.16) | **CONCH (CLIPath, 94.16; +0.02)** |
| NFT Detection | MMRL (98.89) | MMRL (98.67) | MMRL++ (98.97) | MMRL (98.02) | **QuiltNet (MMRL++, 98.97; +0.08)** |
| NCT Classification | MMRL++ (99.45) | CLIPath (99.54) | MMRL++ (99.79) | MMRL++ (99.67) | **QuiltNet (MMRL++, 99.79; +0.12)** |

## 2 Methods

**PEFT Taxonomy:** We structure PEFT methods along three axes: ***Axis A: Adapted Modality (What).*** *Text-only* (learn prompts; freeze vision), *image-only* (vision-side updates; e.g., ProVP, linear probe, vision-LoRA), and *multimodal* (joint text–image adaptation; e.g., MaPLe, MMRL); ***Axis B: Parameterization Strategy (How) -*** *Prompting* (static/conditional), *adapters* (bottlenecks/residual fusion), *LoRA* (low-rank updates), and *mixed*; ***Axis C: Adaptation Locus (Where).*** *Input-level*,*all layers*, *shallow* (early layers), *deep* (late semantic layers), and *embedding/head-level*; Table A13 in Appendix described the method divided by axes in more details.

**Datasets:** We use three datasets in our benchmark experiments. **1) NCT-CRC-HE** Kather et al. [2019] contains H&E stained 100K training and 7,180 test set patches labeled with 9 types of **colorectal cancer** tissues. The distinctive textures make it a comparatively easier multi-class benchmark. **2) SICAPv2** [Silva-Rodríguez et al., 2020] consists of H&E stained 18,783 images of fine-grained **prostate cancer** Gleason grading dataset with subtle glandular and nuclear cues. Appendix Fig. A4 shows the finer morphological cues making it a relatively difficult classification task. **3) NFT:** We curated a novel neuropathology dataset, consisting of ∼3,961 images of neurofibrillary tangles (NFTs) from datasets by [Ghandian et al., 2024] and [Vizcarra et al., 2023]. The Emory dataset, stained using PHF-1 antibody, was scanned at a resolution of 0.23/0.25 microns per pixel by Aperio AT2 scanner; while the UC Davis dataset, stained using AT8 antibody, was scanned at 0.23/0.11 microns per pixel by Zeiss Axio Scan Z1 scanner. The original datasets were point annotated for NFT and preNFTs, with image processing techniques used to obtain NFT object-localization bounding boxes. From these, we sampled 2002 NFT-positive and 1959 control images. **Each image was re-annotated by at least two human annotators, using bounding boxes, to quantify inter-rater variability, reflecting the intrinsic challenges of neuropathology where even specialists disagree on subtle pre-NFT and NFT boundaries** as illustrated in Fig. 2. This process produced a reliability- and localization-aware NFT dataset, containing 1,002 images annotated with additional 1,208 NFTs

Table 2: Classification accuracy and AUCROC comparing CLIP vs CONCH across SICAPv2, NFT, and NCT. The table shows results grouped by adaptation type (text-only, vision-only, multimodal) and reports tuned parameters, compute and memory footprints for CLIP.

| **Method** | **Type** | **Location** | **Tuned params** | **GFLOPs** | **Peak VRAM** | **Model** | **SICAPv2** | | **NFT** | | **NCT** | | **Average** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Acc | AUC | Acc | AUC | Acc | AUC | Acc | AUC |
| *Text-only* | | | | | | | | | | | | | | |
| CoOp | Prompt | *input* level | 0.008M | 124.0 | 0.26 | CLIP | 64.84 | 86.81 | 90.52 | 96.81 | 91.21 | 99.34 | 82.19 | 94.32 |
| | | | | | | CONCH | 59.80 | 82.74 | 90.52 | 96.81 | 88.97 | 98.58 | 79.76 | 92.71 |
| CoCoOp | Prompt | *input* level | 0.035M | 124.0 | 0.26 | CLIP | 62.39 | 85.36 | 89.44 | 96.27 | 89.57 | 99.25 | 80.47 | 93.63 |
| | | | | | | CONCH | 63.24 | 85.02 | 90.92 | 96.48 | 93.05 | 99.49 | 82.40 | 93.66 |
| *Vision-only* | | | | | | | | | | | | | | |
| CLIPath (RFC) | Adapter | *embed. level* | 0.145M | 47.0 | 0.58 | CLIP | 69.84 | 89.13 | 89.64 | 97.44 | 88.14 | 99.41 | 82.54 | 95.33 |
| | | | | | | CONCH | 79.68 | 94.16 | 92.45 | 97.79 | 93.29 | 99.62 | 88.47 | 97.19 |
| $\text{LoRA}_{\text{Img}}$ | LoRA | *all layers* | 0.490M | 35.0 | 0.57 | CLIP | 61.07 | 84.83 | 82.76 | 90.77 | 92.09 | 99.54 | 78.64 | 91.71 |
| | | | | | | CONCH | 71.16 | 91.32 | 79.40 | 93.39 | 95.11 | 99.50 | 81.89 | 94.74 |
| Linear Probe | Adapter | *embed. level* | 0.002M | 35.2 | 0.35 | CLIP | 63.76 | 86.35 | 90.93 | 96.54 | 88.69 | 98.89 | 81.13 | 93.93 |
| | | | | | | CONCH | 79.45 | 93.80 | 92.71 | 97.91 | 95.47 | 99.27 | 89.21 | 96.99 |
| ProVP | Prompt | *all layers* | 0.460M | 88.9 | 0.91 | CLIP | 73.93 | 91.47 | 93.31 | 97.43 | 92.84 | 96.11 | 86.69 | 95.00 |
| | | | | | | CONCH | 64.37 | 84.51 | 93.37 | 97.68 | 94.43 | 98.01 | 84.06 | 93.40 |
| *Multi-modal* | | | | | | | | | | | | | | |
| MaPLe | Prompt | *shallow repr.* | 3.555M | 124.7 | 0.27 | CLIP | 64.93 | 89.50 | 92.19 | 97.24 | 94.15 | 97.85 | 83.76 | 94.86 |
| | | | | | | CONCH | 66.87 | 90.23 | 92.19 | 97.24 | 93.53 | 99.60 | 84.20 | 95.69 |
| MMA | Adapter | *deep repr.* | 0.675M | 70.5 | 0.58 | CLIP | 63.24 | 88.20 | 86.43 | 97.91 | 71.97 | 94.89 | 73.88 | 93.67 |
| | | | | | | CONCH | 43.70 | 71.71 | 49.97 | 65.46 | 23.23 | 68.02 | 38.97 | 68.40 |
| PromptSRC | Prompt | *all layers* | 0.046M | 125.0 | 0.43 | CLIP | 73.89 | 92.27 | 74.55 | 94.20 | 92.24 | 99.13 | 80.23 | 95.20 |
| | | | | | | CONCH | 52.92 | 80.03 | 80.82 | 94.22 | 93.38 | 99.65 | 75.71 | 91.30 |
| MMRL | Mixed | *deep repr.* | 4.992M | 83.9 | 0.87 | CLIP | 75.97 | 92.73 | 95.36 | 98.89 | 96.23 | 98.76 | 89.19 | 96.79 |
| | | | | | | CONCH | 70.88 | 87.06 | 94.34 | 98.02 | 95.50 | 99.17 | 86.91 | 94.75 |
| MMRL++ | Mixed | *deep repr.* | 0.813M | 71.0 | 0.80 | CLIP | 76.53 | 92.28 | 95.31 | 98.86 | 94.11 | 99.45 | 88.65 | 96.86 |
| | | | | | | CONCH | 67.48 | 86.00 | 90.87 | 95.92 | 93.98 | 99.67 | 84.11 | 93.86 |

under strict consensus and 3,614 under union consensus, spanning multiple brain regions, centers, scanners, and antibodies, suitable for evaluating alignment, calibration, and robustness of PEFT methods. Please see more details in Appdendix B.1.

**Evaluation:** We consider two natural image VLMs (CLIP and DFN) and two PFMs (QuiltNet and CONCH) in our benchmark experiments. [2] We selected a set of representative PEFT methods that fine-tune parameters learned over text modality (CoOp, CoCoOp), vision modality (CLIP-RFC, ProVP, Linear Probing) and multi modality (MaPLe, MMA, MMRL, MMRL++) settings. These methods are further categorized based on their parameterization strategy type (e.g., linear probing vs. prompt-based) and location (input level, shallow, vs. deep) in Table 1. Details about the experimental setups and the models/methods are included in Appendix B.2 and C. We applied these PEFT methods to each of the four VLMs/PFM models for the classification tasks using the three pathology datasets described above. We evaluate the efficacy of PEFT techniques under two data availability regimes. (1) **Full data:** We use PEFT to finetune the VLM/PFM on 100% of the training set of the pathology target datasets. Performance metrics like accuracy, AUCROC, and efficiency metrics like GFLOPs and VRAM utilization are recorded. For each task, we select the PEFT method scoring the highest AUCROC score as the best/winning method for the VLM/PFM. **(2) Data efficiency:** We finetune four models in a few shot training setup using PEFT. We randomly select 1, 8 and 16 labeled images of each class from the target datasets and report average accuracies (repeated experiments with three random seeds for each method.We visualize the explanation maps by the PEFT finetuned VLM for a subset of true positive predicted samples from NFT detection task. Explanation maps are generated by using gradcam over the last or second last attention layer of the vision transfomer block of the VLM. Explanation maps generated by the zero shot VLM for the same samples are provided as a baseline to illustrate the human-AI task alignment capabilities of the best PEFT methods.

## 3 Results

Fig. 1 shows CONCH achieves much higher zero shot classification accuracy for pathology datasets than CIFAR10 because its pretraining corpora is more aligned with the pathology domain than natural images. The reverse is true for CLIP and DFN. QuiltNet's performance is subpar. Nevertheless, under zero-shot setting, all four VLMs (including CLIP and DFN) achieve comparable best performances with PEFT, as shown in Table 1 that lists the specific PEFT methods that achieve the best classification accuracy when applied to the four VLMs of each of the three target pathology datasets and tasks. **This demonstrates that PEFT techniques can successfully bridge domain gap between pretrained datasets (natural images) and target pathology domain.** Due to space limitation, we will focus on comparing CLIP and CONCH for subsequent analysis. Details in Appendix Table A2 and Table A3.

[2] Full code will be available at `https://www.github.com/ucdrubinet/VLM_PEFT`

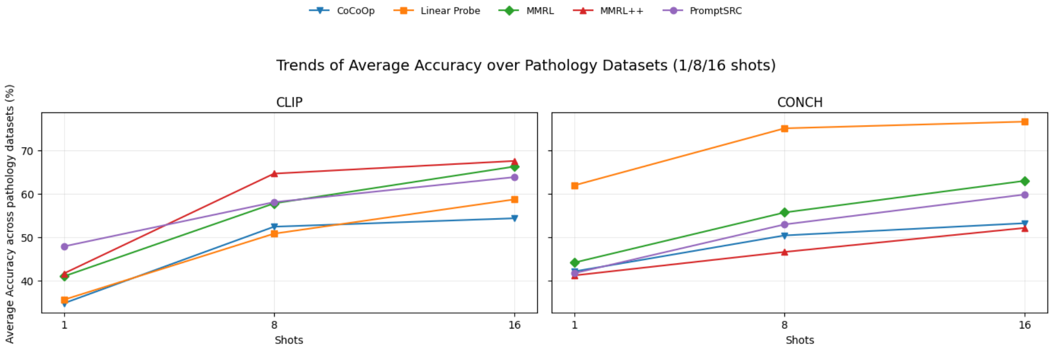

Figure 3: Avg. few-shot accuracies of PEFT methods (%) on 3 pathology tasks for CLIP & CONCH.

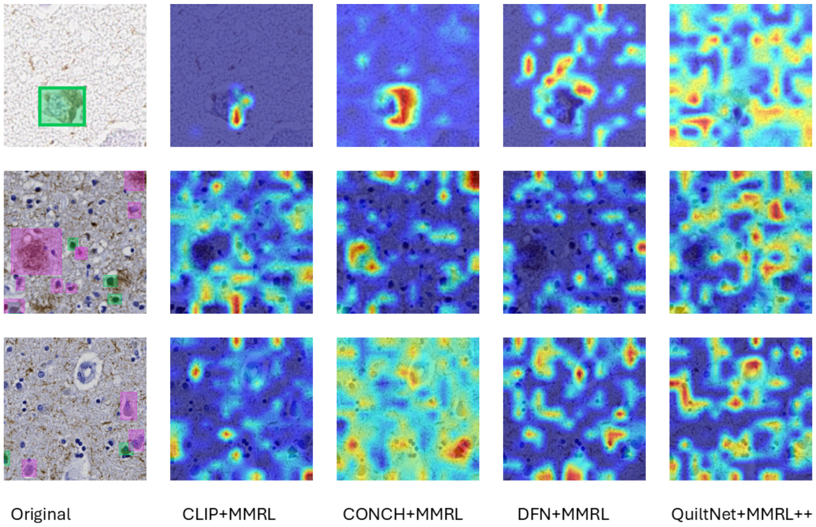

Figure 4: VLM-PEFT task alignment capability across stains : Each row shows an original input tile (first column), AT8-stained in the first row and PHF-1–stained in the remaining rows, followed by Grad-CAM heatmaps from four VLMs fine-tuned with their best performing PEFT method. Consensus NFT regions (NFTs independently identified by both annotators) are outlined with green bounding boxes, while singly annotated regions (NFTs marked by only one annotator) are shown with purple bounding boxes in the first column. Heatmaps use a cold→hot colormap (blue = low, red = high) to indicate the areas most relied upon for NFT prediction. In single-NFT tiles, CLIP+MMRL and CONCH+MMRL concentrate activations within green boxes, whereas in multi-NFT tiles their focus scatters or drifts toward purple regions, revealing reduced reliability under annotation ambiguity

**(i) Full-data adaptation:** For CLIP (a natural image VLM), along **Axis A-Modality**, *multimodal* methods dominate, with MMRL/MMRL++ achieving the best performance across SICAPv2, NFT, and NCT datasets, and MaPLe (94.15/97.85) beating all image- or text-only methods on NCT. Comparing different **parameterization strategies (Axis B))** for vision-only PEFTs reveal that prompt based approach (ProVP 0.46M) and adapters (CLIPath 0.145M) are lightweight but effective, compared to LoRA. For multi-modal adaptation, *mixed* modules (MMRL++ and MMRL) achieve the strongest results over prompt-only or adapter-only methods (MaPLe, MMA, and PromptSRC). When we compare the **adaptation location (Axis C)** of prompt-based approaches, *deep/all-layer* parameter updates are decisively superior, especially on morphology-rich SICAPv2 where MMRL++ (76.53 Acc/92.28 AUC) outperform shallow, input, or embedding loci (MaPLe 64.93/89.50, CoCoOp 64.84/86.81, CLIPath 69.84/89.13).

For **CONCH (pathology VLM)**, the advantage of multi-modality (Axis A) is less apparent: *vision-only* methods performs best on SICAPv2 while *multimodal* wins on NFT tasks and NCT shows a tie. Comparing different **parameterization strategies (Axis B))** for vision-only PEFTs, *adapters/linear heads* methods such as CLIPath and Linear Probe excel perform better than prompting or LoRA.

Along **Axis C**, *embedding/head-level* linear adapters suffice for SICAPv2 and NCT, but *deep* updates are needed for NFT (MMRL 94.34/98.02); shallow/input loci trail on morphology-heavy SICAPv2. Overall, the trend highlights a key divergence in VLM and PFM: **CLIP** requires *deep, mixed, multimodal* adaptation to counter domain shift, whereas **CONCH**, already domain-aligned, favors lightweight *image-only head/embedding* updates on morphology-focused tasks, reserving *deep multimodal* adaptation for NFT where cross-modal alignment matters most.

**(ii) Few-shot settings:** (Fig. 3) shows average few shot performance over 3 datasets. We observe a domain-contingent pattern aligned with our taxonomy. **CLIP (natural-image VLM):** *PromptSRC* (prompting, all-layers, multimodal) best exploits priors and wins with 1-shot learning. With 8–16 shots, *deep/mixed multimodal* modules (e.g., MMRL++) overtake, while shallow prompting (MaPLe) and deep adapters (MMA) lag the linear adapter methods. The gains are largest from 1→8 and then taper. **CONCH (pathology VLM):** across 1/8/16 shots, updating *vision-only embeddings* methods such as Linear probe and ClIPath consistently lead in performance. The performance of multimodal adaptation improves with 8-16 shots but saturates early and remains inferior to vision-only methods. This validates our conjecture that pretrained PFM with knowledge of pathology morphology features require only a few thousand tunable parameters. CLIP benefits from *multimodal* (Axis A), *mixed* (Axis B) and *deep* (Axis C) parameter adaptation as available target labels increase. CONCH prefers lightweight *vision-only, embedding-level* parameter tuning to achieve good performance with fewer shots. Overall, pathology-pretrained encoders retain a clear edge in extreme low-label regimes (Appendix Table A4), whereas with modest labels (16-shots), lightweight multimodal PEFT becomes robust and scalable, delivering comparable performance by tuning only a small fraction of weights.

**(iii) Task alignment:** Fig. 4 illustrates these strengths and limitations by comparing model explanations against pathologist annotations. In *single-NFT tiles*, CLIP+MMRL and CONCH+MMRL generate focused high-activation maps that overlap with the consensus (green) boxes, showing that the interpretation of the target task by both VLM/PFM aligns with human annotators in single-NFT cases. In contrast, *multi-NFT tiles* reveal the weaknesses of the models: activations scatter across the tissue or shift to siloed annotated regions (purple), underscoring their sensitivity to task understanding in complex cases with annotation variability. CONCH+MMRL overlaps more often with at least one annotation mark, suggesting that domain-specific priors help, but alignment remains inconsistent.

## 4 Conclusion

Our benchmarking experiments demonstrate that PEFT is a practical alternative to pretraining large VLMs with target pathology data. With modest labeled data, mixed multimodal and deeper PEFTs can close the domain/lexicon gap and enable generalist VLMs (e.g., CLIP) to match or exceed PFMs (e.g.,CONCH). Under extreme few-shot conditions, CONCH retain an advantage, though multimodal PEFTs close the performance gap with increasing number of target labels. Visualization reveals that PEFT improves task alignment in simple, single-NFT cases (CLIP+MMRL, CONCH+MMRL) but fails to consistently localize consensus regions in tiles with multiple or subtle aggregates, exposing limitations under annotation variability. Our work provides a road map for future research at the intersection of performance, data/model efficiency, and interpretability, helping to bridge the gap between VLMs and practical deployment in real-world pathology settings.

## Acknowledgements

This work was supported in part by funding [2024-351073] from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. Further funding support, in part, was provided by the National Institute on Aging (NIA) of the National Institutes of Health (NIH) under Award Numbers R01AG062517, P30AG072972, P30AG062429, P50AG008702, P30AG066462, and U24NS133949. This work used the Delta GPU cluster at NCSA through allocation [CIS240295] from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants 2138259, 2138286, 2138307, 2137603, and 2138296 [Boerner et al., 2023].

# References

Irina Alafuzoff, Thomas Arzberger, Safa Al-Sarraj, Istvan Bodi, Nenad Bogdanovic, Heiko Braak, Orso Bugiani, Kelly Del-Tredici, Isidro Ferrer, Ellen Gelpi, et al. Staging of neurofibrillary pathology in alzheimer's disease: a study of the brainnet europe consortium. *Brain pathology*, 18(4):484–496, 2008.

Rohan Bareja, Francisco Carrillo-Perez, Yuanning Zheng, Marija Pizurica, Tarak Nath Nandi, Jeanne Shen, Ravi Madduri, and Olivier Gevaert. Evaluating vision and pathology foundation models for computational pathology: A comprehensive benchmark study. *medRxiv*, pages 2025–05, 2025.

Timothy J Boerner, Stephen Deems, Thomas R Furlani, Shelley L Knuth, and John Towns. Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and experience in advanced research computing 2023: Computing for the common good*, pages 173–176. 2023.

Dibaloke Chanda, Milan Aryal, Nasim Yahya Soltani, and Masoud Ganji. A new era in computational pathology: A survey on foundation and vision-language models. *arXiv preprint arXiv:2408.14496*, 2024.

Brittany N Dugger and Dennis W Dickson. Pathology of neurodegenerative diseases. *Cold Spring Harbor perspectives in biology*, 9(7):a028035, 2017.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

Sina Ghandian, Liane Albarghouthi, Kiana Nava, Shivam R Rai Sharma, Lise Minaud, Laurel Beckett, Naomi Saito, Charles DeCarli, Robert A Rissman, Andrew F Teich, et al. Learning precise segmentation of neurofibrillary tangles from rapid manual point annotations. *bioRxiv*, 2024.

Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25015–25025, 2025a.

Yuncheng Guo and Xiaodong Gu. Mmrl++: Parameter-efficient and interaction-aware representation learning for vision-language models. *arXiv preprint arXiv:2505.10088*, 2025b.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

Andrey Ignatov and Grigory Malivenko. Nct-crc-he: Not all histopathological datasets are equally useful. In *European Conference on Computer Vision*, pages 300–317. Springer, 2024.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017, 2023.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16 (1):e1002730, 2019.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023.

Zhengfeng Lai, Zhuoheng Li, Luca Cerny Oliveira, Joohi Chauhan, Brittany N Dugger, and Chen-Nee Chuah. Clipath: Fine-tune clip with visual feature fusion for pathology image analysis towards minimizing data collection efforts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2374–2380, 2023.

Zhengfeng Lai, Joohi Chauhan, Brittany N Dugger, and Chen-Nee Chuah. Bridging the pathology domain gap: Efficiently adapting clip for pathology image analysis with limited labeled data. In *European Conference on Computer Vision*, pages 256–273. Springer, 2024.

Jaeung Lee, Jeewoo Lim, Keunho Byeon, and Jin Tae Kwak. Benchmarking pathology foundation models: Adaptation strategies and scenarios. *Computers in Biology and Medicine*, 190:110031, 2025.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

Zheda Mai, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Quang-Huy Nguyen, Li Zhang, and Wei-Lun Chao. Lessons and insights from a unifying study of parameter-efficient fine-tuning (peft) in visual recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14845–14857, 2025.

Mieko Ochi, Daisuke Komura, and Shumpei Ishikawa. Pathology foundation models. *JMA journal*, 8(1): 121–130, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Julio Silva-Rodríguez, Adrián Colomer, María A Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer methods and programs in biomedicine*, 195:105637, 2020.

Juan C Vizcarra, Thomas M Pearce, Brittany N Dugger, Michael J Keiser, Marla Gearing, John F Crary, Evan J Kiely, Meaghan Morris, Bartholomew White, Jonathan D Glass, et al. Toward a generalizable machine learning workflow for neurodegenerative disease staging with focus on neurofibrillary tangles. *Acta neuropathologica communications*, 11(1):202, 2023.

Conghao Xiong, Hao Chen, and Joseph JY Sung. A survey of pathology foundation model: Progress and future directions. *arXiv preprint arXiv:2504.04045*, 2025.

Chen Xu, Yuhan Zhu, Haocheng Shen, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. Progressive visual prompt learning with contrastive feature re-formation. *International Journal of Computer Vision*, 133 (2):511–526, 2025.

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837, 2024.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

# A Additional Results

Table A1: Zero-shot classification accuracy (%) of open-domain and pathology-specific VLMs on SICAPv2, NFT, and NCT datasets, with averaged pathology accuracy.

| **Method** | **SICAPv2** | **NFT** | **NCT** | **Avg. Pathology Acc.** |
|---|---|---|---|---|
| *Zero-shot Open VLMs* | | | | |
| CLIP | 31.57 | 64.71 | 23.21 | 39.83 |
| DFN | 40.19 | 67.56 | 34.82 | 47.52 |
| *Zero-shot Pathology VLMs* | | | | |
| CONCH | 40.81 | 67.56 | 61.49 | 56.62 |
| QuiltNet | 25.30 | 50.38 | 26.75 | 34.14 |

Table A2: Classification accuracy and AUCROC of natural-image VLMs (CLIP, DFN) adapted using PEFT methods across SICAPv2, NFT, and NCT. The table stratifies methods into text-only, image-only, and multimodal groups. Results show that text-only methods improve accuracy but often plateau in AUCROC, image-only methods like ProVP improve both, and multimodal adapters (MMRL, MMRL++) achieve the highest balance of accuracy and AUCROC. This demonstrates that accuracy and AUCROC can diverge across modalities, with AUCROC revealing robustness missed by accuracy alone.

| **Method** | **Type** | **Location** | **Model** | **SICAPv2** | | **NFT** | | **NCT** | | **Avg** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Acc | AUC | Acc | AUC | Acc | AUC | Acc | AUC |
| *Text-only* | | | | | | | | | | | |
| CoOp | Prompt | *input* level | CLIP | 64.84 | 86.81 | 90.52 | 96.81 | 91.21 | 99.34 | 82.19 | 94.32 |
| | | | DFN | 68.57 | 88.64 | 90.26 | 96.43 | 92.92 | 99.50 | 83.92 | 94.86 |
| CoCoOp | Prompt | *input* level | CLIP | 62.39 | 85.36 | 89.44 | 96.27 | 89.57 | 99.25 | 80.47 | 93.63 |
| | | | DFN | 64.80 | 75.80 | 90.46 | 96.71 | 92.26 | 99.37 | 82.51 | 90.63 |
| *Vision-only* | | | | | | | | | | | |
| CLIPath (RFC) | Adapter | *embed. level* | CLIP | 69.84 | 89.13 | 89.64 | 97.44 | 88.14 | 99.41 | 82.54 | 95.33 |
| | | | DFN | 69.08 | 88.16 | 91.89 | 97.49 | 92.33 | 99.54 | 84.43 | 95.06 |
| LoRA$_{\text{Img}}$ | LoRA | *all layers* | CLIP | 61.07 | 84.83 | 82.76 | 90.77 | 92.09 | 99.54 | 78.64 | 91.71 |
| | | | DFN | 65.60 | 86.36 | 88.58 | 95.54 | 89.89 | 99.35 | 81.36 | 93.75 |
| Linear Probe | Adapter | *embed. level* | CLIP | 63.76 | 86.35 | 90.93 | 96.54 | 88.69 | 98.89 | 81.13 | 93.93 |
| | | | DFN | 67.58 | 87.51 | 91.59 | 97.12 | 87.70 | 98.58 | 82.29 | 94.40 |
| ProVP | Prompt | *all layers* | CLIP | 73.93 | 91.47 | 93.31 | 97.43 | 92.84 | 96.11 | 86.69 | 95.00 |
| | | | DFN | 74.36 | 92.00 | 93.52 | 97.33 | 92.68 | 99.52 | 86.85 | 96.28 |
| *Multi-modal* | | | | | | | | | | | |
| MaPLe | Prompt | *shallow repr.* | CLIP | 64.93 | 89.50 | 92.19 | 97.24 | 94.15 | 97.85 | 83.76 | 94.86 |
| | | | DFN | 75.58 | 93.14 | 93.52 | 97.79 | 92.30 | 97.88 | 87.13 | 96.27 |
| MMA | Adapter | *deep repr.* | CLIP | 63.24 | 88.20 | 86.43 | 97.91 | 71.97 | 94.89 | 73.88 | 93.67 |
| | | | DFN | 69.50 | 90.26 | 91.89 | 97.62 | 51.61 | 94.79 | 71.00 | 94.22 |
| PromptSRC | Prompt | *all layers* | CLIP | 73.89 | 92.27 | 74.55 | 94.20 | 92.24 | 99.13 | 80.23 | 95.20 |
| | | | DFN | 72.05 | 91.73 | 82.86 | 95.27 | 91.42 | 98.95 | 82.11 | 95.32 |
| MMRL | Mixed | *deep repr.* | CLIP | 75.97 | 92.73 | 95.36 | 98.89 | 96.23 | 98.76 | 89.19 | 96.79 |
| | | | DFN | 76.01 | 92.70 | 94.80 | 98.67 | 96.11 | 99.51 | 88.97 | 96.96 |
| MMRL++ | Mixed | *deep repr.* | CLIP | 76.53 | 92.28 | 95.31 | 98.86 | 94.11 | 99.45 | 88.65 | 96.86 |
| | | | DFN | 76.86 | 92.96 | 95.16 | 98.46 | 95.72 | 99.40 | 89.25 | 96.94 |

Table A3: Classification Accuracy and AUCROC of various Path VLMs + PEFT techniques after training on 100% of SICAPv2, NFT, and NCT datasets, with the adaptation location. An Avg column reports the mean Acc/AUC over the three datasets.

| **Method** | **Type** | **Location** | **Model** | **SICAPv2** | | **NFT** | | **NCT** | | **Avg** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Acc | AUC | Acc | AUC | Acc | AUC | Acc | AUC |
| *Text-only* | | | | | | | | | | | |
| CoOp | Prompt | *input* level | QuiltNet | 74.79 | 92.23 | 91.74 | 96.79 | 93.15 | 98.25 | 86.56 | 95.76 |
| | | | CONCH | 59.80 | 82.74 | 90.52 | 96.81 | 88.97 | 98.58 | 79.76 | 92.71 |
| CoCoOp | Prompt | *input* level | QuiltNet | 74.93 | 92.31 | 90.62 | 95.83 | 90.96 | 99.28 | 85.50 | 95.81 |
| | | | CONCH | 63.24 | 85.02 | 90.92 | 96.48 | 93.05 | 99.49 | 82.40 | 93.66 |
| *Vision-only* | | | | | | | | | | | |
| CLIPath (RFC) | Adapter | *embed. level* | QuiltNet | 72.62 | 90.91 | 92.29 | 97.53 | 71.21 | 95.48 | 78.71 | 94.64 |
| | | | CONCH | 79.68 | 94.16 | 92.45 | 97.79 | 93.29 | 99.62 | 88.47 | 97.19 |
| LoRA$_{\text{Img}}$ | LoRA | *all layers* | QuiltNet | 71.35 | 90.19 | 83.27 | 92.28 | 93.62 | 99.69 | 82.75 | 94.05 |
| | | | CONCH | 71.16 | 91.32 | 79.40 | 93.39 | 95.11 | 99.50 | 81.89 | 94.74 |
| Linear Probe | Adapter | *embed. level* | QuiltNet | 68.19 | 90.92 | 86.28 | 95.98 | 93.05 | 99.44 | 82.51 | 95.45 |
| | | | CONCH | 79.45 | 93.80 | 92.71 | 97.91 | 95.47 | 99.27 | 89.21 | 96.99 |
| ProVP | Prompt | *all layers* | QuiltNet | 74.00 | 91.42 | 94.79 | 97.53 | 92.47 | 99.38 | 87.09 | 96.11 |
| | | | CONCH | 64.37 | 84.51 | 93.37 | 97.68 | 94.43 | 98.01 | 84.06 | 93.40 |
| *Multi-modal* | | | | | | | | | | | |
| MaPLe | Prompt | *shallow repr.* | QuiltNet | 78.27 | 94.14 | 94.59 | 98.11 | 93.70 | 99.76 | 88.85 | 97.34 |
| | | | CONCH | 66.87 | 90.23 | 92.19 | 97.24 | 93.53 | 99.60 | 84.20 | 95.69 |
| MMA | Adapter | *deep repr.* | QuiltNet | 72.43 | 90.15 | 89.49 | 97.06 | 60.77 | 94.63 | 74.23 | 93.95 |
| | | | CONCH | 43.70 | 71.71 | 49.97 | 65.46 | 23.23 | 68.02 | 38.97 | 68.40 |
| PromptSRC | Prompt | *all layers* | QuiltNet | 75.49 | 93.05 | 76.74 | 93.20 | 94.05 | 99.45 | 82.09 | 95.23 |
| | | | CONCH | 52.92 | 80.03 | 80.82 | 94.22 | 93.38 | 99.65 | 75.71 | 91.30 |
| MMRL | Mixed | *deep repr.* | QuiltNet | 77.95 | 90.15 | 95.26 | 98.80 | 94.80 | 99.76 | 89.34 | 96.24 |
| | | | CONCH | 70.88 | 87.06 | 94.34 | 98.02 | 95.50 | 99.17 | 86.91 | 94.75 |
| MMRL++ | Mixed | *deep repr.* | QuiltNet | 77.76 | 89.81 | 95.87 | 98.97 | 95.54 | 99.79 | 89.72 | 96.19 |
| | | | CONCH | 67.48 | 86.00 | 90.87 | 95.92 | 93.98 | 99.67 | 84.11 | 93.86 |

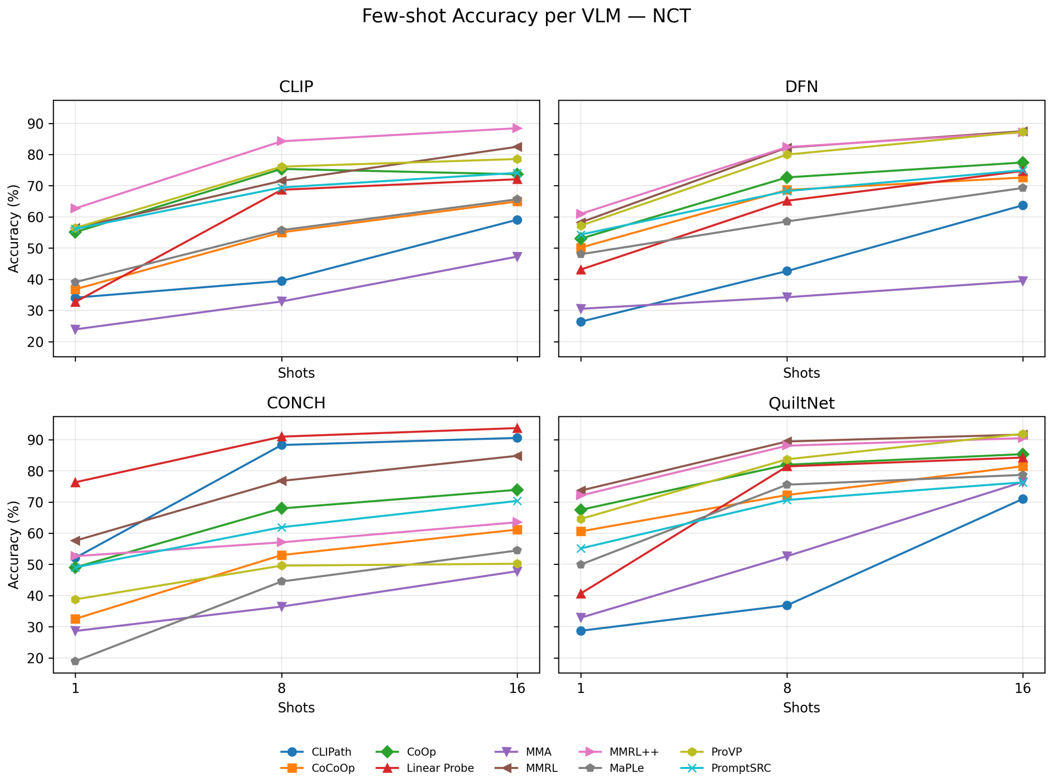

Figure A1: **NCT dataset: few-shot patch classification.** Comparison of PEFT strategies—applied to open-domain CLIP-like VLMs and pathology-pretrained VLMs across $K$-shot regimes ($K \in \{1, 8, 16\}$). Curves summarize the accuracy metric.

## A.1 Taxonomy-based guidance on results

### A.1.1 Full-data Adaptation

**Full-data Adaptation of Natural VLMs (CLIP).** CLIP's behavior follows the taxonomy closely. When *Axis A = text-only*, *Axis B = prompting* (CoOp/CoCoOp), and *Axis C = input-level*, it saturates the coarse NCT task (CoOp: NCT AUC 99.34, Acc 91.21; CoCoOp: AUC 99.25, Acc 89.57) but underperforms on fine-grained tissue labeling (SICAPv2: CoOp Acc 64.84 / AUC 86.81; CoCoOp Acc 62.39 / AUC 85.36). Moving to *Axis A = image-only* with *Axis B = adapters/prompt-like residuals* at *Axis C = embedding/head or all layers* (e.g., CLIPath at embedding-level; ProVP across all layers) yields steadier gains across datasets (CLIPath: SICAPv2 Acc 69.84 / AUC 89.13; ProVP: NFT Acc 93.31 / AUC 97.43). Finally, compact *Axis A = multimodal* methods with *Axis B = mixed* at *Axis C = deep representations* (MMRL/MMRL++) deliver the best overall balance (MMRL: SICAPv2 Acc 75.97 / AUC 92.73; NFT Acc 95.36 / AUC 98.89; MMRL++: NCT AUC 99.45 with 0.813M tuned params), indicating that cross-modal, deep-locus adaptation is especially helpful for rare-lesion and fine-grained settings. Appendix A2 details these trends (and DFN).

**Full-data Adaptation of Pathology VLMs (CONCH).** CONCH's strongest gains appear under *Axis A = image-only* with *Axis B = adapters/linear readout* at *Axis C = embedding/head-level*. Vision-focused updates (CLIPath, Linear Probe) produce the largest improvements on fine-grained SICAPv2 and remain strong on NFT/NCT (CLIPath: SICAPv2 Acc 79.68 / AUC 94.16; NFT Acc 92.45 / AUC 97.79; NCT AUC 99.62; Linear Probe: SICAPv2 Acc 79.45 / AUC 93.80; NFT Acc 92.71 / AUC 97.91; NCT Acc 95.47 / AUC 99.27). Text-only prompting (*A=text-only*, *B=prompting*, *C=input*) still achieves NCT ceilings (CoOp NCT AUC 98.58; CoCoOp AUC 99.49) but is weak on SICAPv2 (CoOp Acc 59.80 / AUC 82.74). Multimodal mixed strategies at deep layers (MMRL: NFT Acc 94.34 / AUC 98.02) are competitive on NFT/NCT but show variable gains versus vision-only

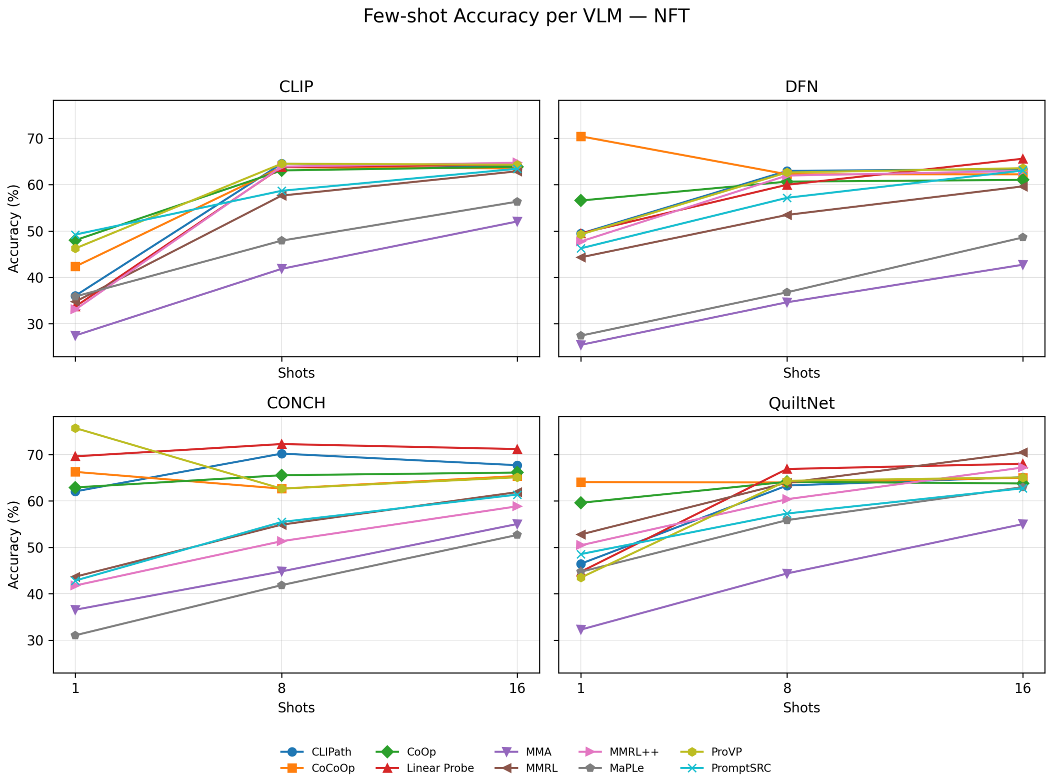

Figure A2: **NFT dataset: few-shot patch classification.** Comparison of PEFT strategies—applied to open-domain CLIP-like VLMs and pathology-pretrained VLMs across $K$-shot regimes ($K \in \{1, 8, 16\}$). Curves summarize the accuracy metric.

adapters—consistent with CONCH's pathology-specific pretraining favoring lightweight vision-side adaptation. Appendix A3 expands on these patterns alongside QuiltNet.

**Natural vs. Pathology VLMs.** Task winners align more with taxonomy choices than pretraining domain alone. For SICAPv2 (fine-grained), the winning recipe is *Axis A* = *image-only* + *Axis C* = *embedding/head (shallow)* with *Axis B* = *adapters/linear*: CONCH+CLIPath (Acc 79.68 / AUC 94.16) and CONCH+Linear Probe (Acc 79.45 / AUC 93.80) outperform CLIP's best text-only prompts. For NFT, higher-capacity *Axis A* = *multimodal* with *Axis B* = *mixed* at *Axis C* = *deep repr.* slightly favors CLIP at the top end (CLIP+MMRL: Acc 95.36 / AUC 98.89; CLIP+MMRL++: Acc 95.31 / AUC 98.86), while CONCH remains competitive under vision-centered adapters (CLIPath: Acc 92.45 / AUC 97.79). For NCT (coarse), *Axis A* = *text-only*, *Axis B* = *prompting*, *Axis C* = *input* attains ∼99% AUC for both families (CLIP CoOp/CoCoOp: 99.34/99.25; CONCH CoOp/CoCoOp: 98.58/99.49), so the edge depends on minor PEFT choices and whether one optimizes Acc vs. AUC.

# B Experiment Setup

## B.1 Datasets

In our benchmark experiments, we evaluate three histopathology datasets spanning colorectal cancer, prostate cancer, and neuropathology.

**NCT-CRC-HE** is a large-scale colorectal cancer dataset derived from hematoxylin & eosin (H&E)–stained whole slide images collected from the National Center for Tumor Diseases (NCT, Heidelberg) and the University Medical Center Mannheim (UMM), Germany Kather et al. [2019]. It contains 100,000 training and 7,180 independent test patches, the latter commonly referred to as CRC-VAL-HE-7K. Each patch is 224 × 224 pixels at 0.5 µm per pixel (∼20× magnification), and all images were color normalized to mitigate staining variability Kather et al. [2019]. Patches are anno-

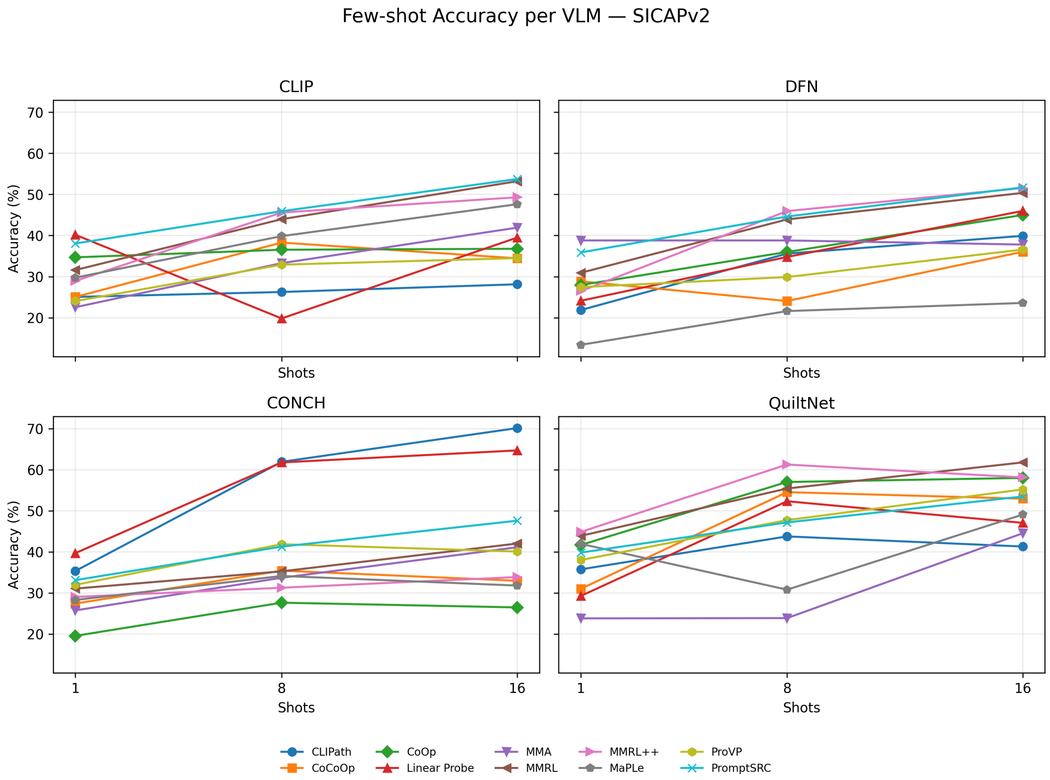

Figure A3: **Sicapv2 dataset: few-shot patch classification.** Comparison of PEFT strategies—applied to open-domain CLIP-like VLMs and pathology-pretrained VLMs across $K$-shot regimes ($K \in \{1, 8, 16\}$). Curves summarize the accuracy metric.

tated across nine histologic classes, including normal mucosa, tumor epithelium, cancer-associated stroma, smooth muscle, lymphocytes, debris, mucus, adipose tissue, and background. Owing to its size and standardized train/test split, it has become a widely used benchmark for tissue classification; however, the dataset is comparatively easier, as models can sometimes exploit low-level color or compression artifacts rather than robust morphological cues Ignatov and Malivenko [2024].

**SICAPv2** is a prostate cancer dataset designed for Gleason grading and fine-grained morphological analysis. It consists of 18,783 image patches of size $512 \times 512$ pixels at $10\times$ magnification, sampled from 155 annotated whole slide images of biopsies and prostatectomy specimens Silva-Rodríguez et al. [2020]. Unlike texture-driven datasets such as NCT-CRC-HE, prostate grading requires recognition of subtle glandular architecture and nuclear morphology, which is intrinsically more challenging.

**NFT** is a dataset designed for detecting neurofibrillary tangles, a hallmark for Alzheimer's disease. We sampled 3961 images randomly from these two datasets. To ensure reliability, we re-annotated the NFT dataset with two independent annotators. The test set of the dataset contained 1,961 images, each with a single NFT annotation from a previous annotator. Our QC process revealed that many images contained additional NFTs and that annotators frequently disagreed on NFT boundaries. Annotator group G1 marked 2,571 NFT boxes (mean $\sim$1.10 per image, max 8), while Annotator group G2 marked 2,320 NFT boxes (mean $\sim$1.04 per image, max 9). From these, an intersection consensus (NFTs agreed upon by both annotators) produced 1,208 NFTs across 1,173 images (mean $\sim$1.03 per image, max 5). In contrast, a union consensus (NFTs marked by either annotator) expanded to 3,614 NFTs across 2,340 images (mean $\sim$1.54 per image, max 13). This QC process transformed the original single-annotation dataset into a reliability- and localization-aware NFT resource, explicitly capturing inter-rater variability and the true multiplicity of tangles per image, as illustrated in Figure 2. Together, this yields a reliability- and localization-aware NFT corpus with explicit rater variability, heterogeneous antibodies (PHF-1, pan-tau, AT8, CP13), multiple centers and scanners (Aperio AT2, ZEISS AxioScan), sub-micron pixel resolutions (0.11, 0.23–0.25 µm/pixel), and varied

Table A4: Few-shot classification accuracy (%) of PEFT techniques on SICAPv2, NFT, and NCT grouped by method with subrows for the underlying VLM (CLIP, DFN, QuiltNet, CONCH). Each dataset shows 1-shot (1S), 8-shot (8S), and 16-shot (16S) accuracy; **Avg** columns are per-shot means across datasets (SICAPv2, NFT, NCT). We dont showcase results from LoRA$_{\text{Img}}$ because of failures in full data adaptation

| **Method** | **Adaptation Type** | **Adaptation Location** | **Model** | **SICAPv2** | | | **NFT** | | | **NCT** | | | **Avg** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 1S | 8S | 16S | 1S | 8S | 16S | 1S | 8S | 16S | 1S | 8S | 16S |
| *Adapted modality: Text* | | | | | | | | | | | | | | | |
| CoOp | Prompt | Input-level | CLIP | 34.69 | 36.54 | 36.77 | 48.02 | 63.06 | 63.84 | 55.13 | 75.38 | 73.67 | 45.95 | 58.33 | 58.09 |
| | | | DFN | 27.95 | 36.00 | 45.00 | 56.57 | 60.65 | 61.02 | 53.03 | 72.66 | 77.44 | 45.85 | 56.44 | 61.15 |
| | | | QuiltNet | 41.67 | 57.01 | 58.00 | 59.60 | 64.13 | 63.76 | 67.49 | 81.97 | 85.35 | 56.25 | 67.70 | 69.04 |
| | | | CONCH | 19.54 | 27.62 | 26.47 | 62.91 | 65.54 | 66.12 | 49.04 | 67.95 | 73.92 | 43.83 | 53.70 | 55.50 |
| CoCoOp | Prompt | Input-level | CLIP | 25.04 | 38.28 | 34.45 | 42.34 | 63.75 | 63.50 | 36.79 | 55.04 | 64.94 | 34.72 | 52.36 | 54.30 |
| | | | DFN | 28.84 | 24.05 | 36.04 | 70.41 | 62.23 | 62.22 | 50.15 | 68.65 | 72.61 | 49.80 | 51.64 | 56.96 |
| | | | QuiltNet | 30.99 | 54.54 | 52.94 | 64.06 | 63.98 | 65.00 | 60.59 | 72.25 | 81.46 | 51.88 | 63.59 | 66.47 |
| | | | CONCH | 27.41 | 35.42 | 32.96 | 66.26 | 62.65 | 65.36 | 32.54 | 52.99 | 61.17 | 42.07 | 50.35 | 53.16 |
| *Adapted modality: Image* | | | | | | | | | | | | | | | |
| Linear Probe | Adapter | Embed. / head-level | CLIP | 40.18 | 19.81 | 39.51 | 33.72 | 63.76 | 64.32 | 32.72 | 68.69 | 72.08 | 35.54 | 50.75 | 58.64 |
| | | | DFN | 24.11 | 34.78 | 45.99 | 49.50 | 59.97 | 65.58 | 43.11 | 65.17 | 74.74 | 38.91 | 53.31 | 62.10 |
| | | | QuiltNet | 29.26 | 52.32 | 47.06 | 44.74 | 66.89 | 68.01 | 40.66 | 81.43 | 84.26 | 38.22 | 66.88 | 66.44 |
| | | | CONCH | 39.68 | 61.77 | 64.70 | 69.61 | 72.26 | 71.19 | 76.34 | 90.97 | 93.72 | 61.88 | 75.00 | 76.54 |
| CLIPath (RFC) | Adapter | Embed. / head-level | CLIP | 25.04 | 26.25 | 28.13 | 36.05 | 64.52 | 63.86 | 34.10 | 39.46 | 59.10 | 31.73 | 43.41 | 50.36 |
| | | | DFN | 21.85 | 35.60 | 39.92 | 49.45 | 62.96 | 63.35 | 26.40 | 42.63 | 63.72 | 32.57 | 47.06 | 55.66 |
| | | | QuiltNet | 35.74 | 43.76 | 41.30 | 46.46 | 63.32 | 65.12 | 28.71 | 36.87 | 70.99 | 36.97 | 47.98 | 59.14 |
| | | | CONCH | 35.34 | 61.92 | 70.14 | 62.06 | 70.20 | 67.70 | 52.11 | 88.28 | 90.57 | 49.84 | 73.47 | 76.14 |
| ProVP | Prompt | All layers | CLIP | 24.03 | 32.90 | 34.50 | 46.25 | 64.51 | 64.29 | 56.47 | 76.11 | 78.56 | 42.25 | 57.84 | 59.12 |
| | | | DFN | 27.43 | 29.89 | 36.58 | 49.26 | 62.62 | 63.54 | 57.21 | 79.97 | 87.29 | 44.63 | 57.49 | 62.47 |
| | | | QuiltNet | 38.00 | 47.73 | 55.20 | 43.48 | 64.34 | 65.07 | 64.57 | 83.70 | 91.88 | 48.68 | 65.26 | 70.72 |
| | | | CONCH | 31.97 | 41.86 | 40.07 | 75.71 | 62.64 | 65.10 | 38.78 | 49.60 | 50.21 | 48.82 | 51.37 | 51.79 |
| *Adapted modality: Multi-modal* | | | | | | | | | | | | | | | |
| PromptSRC | Prompt | all-layers | CLIP | 38.06 | 45.92 | 53.76 | 49.23 | 58.67 | 63.42 | 56.17 | 69.44 | 74.12 | 47.82 | 58.01 | 63.77 |
| | | | DFN | 35.88 | 44.62 | 51.73 | 46.27 | 57.16 | 63.04 | 54.32 | 68.28 | 74.95 | 45.49 | 56.69 | 63.24 |
| | | | QuiltNet | 39.82 | 47.16 | 53.48 | 48.54 | 57.29 | 62.73 | 55.12 | 70.65 | 76.31 | 47.83 | 58.37 | 64.17 |
| | | | CONCH | 33.11 | 41.27 | 47.59 | 42.83 | 55.46 | 61.37 | 49.05 | 61.92 | 70.36 | 41.66 | 52.88 | 59.77 |
| MaPLe | Prompt | Shallow repr | CLIP | 29.74 | 39.85 | 47.63 | 35.88 | 47.92 | 56.34 | 39.07 | 55.75 | 65.65 | 34.90 | 47.84 | 56.54 |
| | | | DFN | 13.38 | 21.60 | 23.58 | 27.44 | 36.79 | 48.61 | 48.02 | 58.54 | 69.27 | 29.61 | 38.98 | 47.15 |
| | | | QuiltNet | 41.95 | 30.79 | 49.08 | 44.72 | 55.86 | 62.98 | 49.95 | 75.56 | 78.70 | 45.54 | 54.07 | 63.59 |
| | | | CONCH | 28.28 | 34.13 | 31.79 | 31.04 | 41.85 | 52.67 | 18.96 | 44.52 | 54.46 | 26.09 | 40.17 | 46.31 |
| MMA | Adapter | Deep repr | CLIP | 22.54 | 33.27 | 41.92 | 27.48 | 41.85 | 52.06 | 23.93 | 32.87 | 47.27 | 24.65 | 36.00 | 47.08 |
| | | | DFN | 38.80 | 38.80 | 37.80 | 25.46 | 34.64 | 42.73 | 30.53 | 34.23 | 39.43 | 31.60 | 35.89 | 39.99 |
| | | | QuiltNet | 23.80 | 23.87 | 44.53 | 32.26 | 44.37 | 55.01 | 32.90 | 52.63 | 76.43 | 29.65 | 40.29 | 58.66 |
| | | | CONCH | 25.74 | 33.69 | 41.12 | 36.55 | 44.80 | 55.03 | 28.64 | 36.47 | 47.82 | 30.31 | 38.32 | 47.99 |
| MMRL | Adapter | Deep (late semantic) | CLIP | 31.59 | 43.97 | 53.22 | 34.77 | 57.64 | 62.88 | 56.42 | 71.55 | 82.49 | 40.93 | 57.72 | 66.20 |
| | | | DFN | 30.92 | 43.92 | 50.38 | 44.33 | 53.47 | 59.62 | 58.24 | 82.22 | 87.47 | 44.50 | 59.87 | 65.82 |
| | | | QuiltNet | 43.89 | 55.42 | 61.78 | 52.77 | 63.92 | 70.46 | 73.68 | 89.42 | 91.64 | 56.78 | 69.59 | 74.63 |
| | | | CONCH | 31.05 | 35.25 | 42.03 | 43.67 | 54.88 | 61.91 | 57.60 | 76.78 | 84.82 | 44.11 | 55.64 | 62.92 |
| MMRL++ | Adapter | Deep repr | CLIP | 28.95 | 45.59 | 49.31 | 33.07 | 63.89 | 64.72 | 62.73 | 84.25 | 88.44 | 41.58 | 64.58 | 67.49 |
| | | | DFN | 26.46 | 45.95 | 51.57 | 47.75 | 61.95 | 62.95 | 60.96 | 82.38 | 87.10 | 45.06 | 63.43 | 67.21 |
| | | | QuiltNet | 44.85 | 61.26 | 58.13 | 50.44 | 60.37 | 67.21 | 72.10 | 88.06 | 90.49 | 55.80 | 69.90 | 71.94 |
| | | | CONCH | 29.03 | 31.26 | 33.87 | 41.77 | 51.32 | 58.84 | 52.64 | 57.07 | 63.52 | 41.15 | 46.55 | 52.08 |

case demographics — key for evaluating detection, calibration, and cross-site generalization in NFT-pathology modeling.

## B.2 Experimental Evaluation

### B.2.1 Zero Shot

Prompts such as, "a pathology tissue showing", "a photo of" and "a histopathological image of" is appended before the class labels of the respective pathology and natural image datasets, as a prompt tuning method. No training data is provided, and the prompt tuning experiment reporting highest accuracy is recorded for the VLM in Table A1. Figure 1 also illustrates performance on CIFAR-100 dataset to show affinity to natural image vs pathology domain respectively.

### B.2.2 PEFT

We implemented a diverse set of parameter-efficient fine-tuning (PEFT) methods spanning prompt-based, adapter-based, and low-rank approaches in order to evaluate their adaptability across both vision–language models (CLIP, DFN) and pathology foundation models (QuiltNet, CONCH). All experiments were initialized from official pretrained checkpoints, with all parameters frozen except for the lightweight modules introduced by each PEFT method. To ensure rigor and fairness, we closely followed the training protocols recommended in the original works, while conducting narrow hyperparameter tuning grids to address instability or outlier behaviors.

**Text-modality adaptation.** CoOp was trained for 50 epochs with 16 learnable tokens (fp16), initialized with either natural language prompts ("a pathology tissue showing," "a photo of," "a histopathological image of") or pure tokens. CoCoOp used four tokens initialized with three prompts and converged within 10 epochs.

**Image-modality adaptation.** Linear probing trained only a shallow classifier. CLIP-RFC fine-tuned a residual feature connection on ViT-B/16 for 10 epochs with cross-entropy loss, AdamW, learning rate $1 \times 10^{-4}$, and $\alpha = 0.8$. ProVP trained 16 learnable prompts on CLIP ViT-B/16 with SGD, cosine LR scheduler, batch size 64, weight decay 0.0005, and max 50 epochs.

**Multi-modality adaptation.** PromptSRC added both text and vision prompts (4 tokens each, depth 9), with losses weighted following the original paper. MaPLe, MMA, MMRL, and MMRL++ jointly tuned text and vision modules with default bottleneck or prompt dimensions, applying light tuning for CONCH. LoRA inserted rank-16, $\alpha = 32$ low-rank adapters into ViT-B/16 attention layers, trained for 10 epochs with batch size 16, learning rate $1 \times 10^{-4}$, and weight decay 0.005.

All experiments were capped at 10–50 epochs or early-stopped by non-decreasing training loss with a patience of 5 epochs. Few-shot experiments used 1, 8, and 16 samples per class across three seeds; no extra labeled data was provided. FLOPs, trainable parameters, and VRAM usage were recorded on Tesla T4 GPUs (16 GB). Most runs were performed on Tesla T4 GPUs (16 GB), with large-scale LoRA and adapters configuration design also tested on NVIDIA L40S GPUs (48 GB).

Table A5: Classification accuracy and AUCROC for **CLIP** across SICAPv2, NFT, and NCT. Results are grouped by adaptation type.

| Method | Type | Location | Model | SICAPv2 | | NFT | | NCT | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Acc | AUC | Acc | AUC | Acc | AUC | Acc | AUC |
| *Text-only* | | | | | | | | | | | |
| CoOp | Prompt | *input* level | CLIP | 64.84 | 86.81 | 90.52 | 96.81 | 91.21 | 99.34 | 82.19 | 94.32 |
| CoCoOp | Prompt | *input* level | CLIP | 62.39 | 85.36 | 89.44 | 96.27 | 89.57 | 99.25 | 80.47 | 93.63 |
| *Vision-only* | | | | | | | | | | | |
| CLIPath (RFC) | Adapter | *embed. level* | CLIP | 69.84 | 89.13 | 89.64 | 97.44 | 88.14 | 99.41 | 82.54 | 95.33 |
| LoRA$_{\text{Img}}$ | LoRA | *all layers* | CLIP | 61.07 | 84.83 | 82.76 | 90.77 | 92.09 | 99.54 | 78.64 | 91.71 |
| Linear Probe | Adapter | *embed. level* | CLIP | 63.76 | 86.35 | 90.93 | 96.54 | 88.69 | 98.89 | 81.13 | 93.93 |
| ProVP | Prompt | *all layers* | CLIP | 73.93 | 91.47 | 93.31 | 97.43 | 92.84 | 96.11 | 86.69 | 95.00 |
| *Multi-modal* | | | | | | | | | | | |
| MaPLe | Prompt | *shallow repr.* | CLIP | 64.93 | 89.50 | 92.19 | 97.24 | 94.15 | 97.85 | 83.76 | 94.86 |
| MMA | Adapter | *deep repr.* | CLIP | 63.24 | 88.20 | 86.43 | 97.91 | 71.97 | 94.89 | 73.88 | 93.67 |
| PromptSRC | Prompt | *all layers* | CLIP | 73.89 | 92.27 | 74.55 | 94.20 | 92.24 | 99.13 | 80.23 | 95.20 |
| MMRL | Mixed | *deep repr.* | CLIP | 75.97 | 92.73 | 95.36 | 98.89 | 96.23 | 98.76 | 89.19 | 96.79 |
| MMRL++ | Mixed | *deep repr.* | CLIP | 76.53 | 92.28 | 95.31 | 98.86 | 94.11 | 99.45 | 88.65 | 96.86 |

Table A6: Few-shot classification accuracy (%) of PEFT techniques on SICAPv2, NFT, and NCT using **CLIP**. Each dataset shows 1-shot (1S), 8-shot (8S), and 16-shot (16S) accuracy; **Avg** columns are per-shot means across datasets.

| Method | Adaptation Type | Adaptation Location | SICAPv2 | | | NFT | | | NCT | | | Avg | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1S | 8S | 16S | 1S | 8S | 16S | 1S | 8S | 16S | 1S | 8S | 16S |
| *Adapted modality: Text* | | | | | | | | | | | | | | |
| CoOp | Prompt | Input-level | 34.69 | 36.54 | 36.77 | 48.02 | 63.06 | 63.84 | 55.13 | 75.38 | 73.67 | 45.95 | 58.33 | 58.09 |
| CoCoOp | Prompt | Input-level | 25.04 | 38.28 | 34.45 | 42.34 | 63.75 | 63.50 | 36.79 | 55.04 | 64.94 | 34.72 | 52.36 | 54.30 |
| *Adapted modality: Image* | | | | | | | | | | | | | | |
| Linear Probe | Adapter | Embed./head-level | 40.18 | 19.81 | 39.51 | 33.72 | 63.76 | 64.32 | 32.72 | 68.69 | 72.08 | 35.54 | 50.75 | 58.64 |
| CLIPath (RFC) | Adapter | Embed./head-level | 25.04 | 26.25 | 28.13 | 36.05 | 64.52 | 63.86 | 34.10 | 39.46 | 59.10 | 31.73 | 43.41 | 50.36 |
| ProVP | Prompt | All layers | 24.03 | 32.90 | 34.50 | 46.25 | 64.51 | 64.29 | 56.47 | 76.11 | 78.56 | 42.25 | 57.84 | 59.12 |
| *Adapted modality: Multi-modal* | | | | | | | | | | | | | | |
| PromptSRC | Prompt | All layers | 38.06 | 45.92 | 53.76 | 49.23 | 58.67 | 63.42 | 56.17 | 69.44 | 74.12 | 47.82 | 58.01 | 63.77 |
| MaPLe | Prompt | Shallow repr | 29.74 | 39.85 | 47.63 | 35.88 | 47.92 | 56.34 | 39.07 | 55.75 | 65.65 | 34.90 | 47.84 | 56.54 |
| MMA | Adapter | Deep repr | 22.54 | 33.27 | 41.92 | 27.48 | 41.85 | 52.06 | 23.93 | 32.87 | 47.27 | 24.65 | 36.00 | 47.08 |
| MMRL | Adapter | Deep (late sem.) | 31.59 | 43.97 | 53.22 | 34.77 | 57.64 | 62.88 | 56.42 | 71.55 | 82.49 | 40.93 | 57.72 | 66.20 |
| MMRL++ | Adapter | Deep repr | 28.95 | 45.59 | 49.31 | 33.07 | 63.89 | 64.72 | 62.73 | 84.25 | 88.44 | 41.58 | 64.58 | 67.49 |

# C Related Background Literature

## C.1 Vision–Language Models (VLMs)

**Contrastive pretraining.** CLIP-style VLMs comprise an image encoder $f_\theta : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^d$, and a text encoder $g_\phi : \mathcal{T} \to \mathbb{R}^d$, trained on paired image–text data $\{(x_i, t_i)\}_{i=1}^{B}$, with a symmetric InfoNCE objective and temperature $\tau$. With L2-normalized embeddings and dot-product similarity,

Table A7: Classification accuracy and AUCROC for **CONCH** across SICAPv2, NFT, and NCT. Results are grouped by adaptation type.

| **Method** | **Type** | **Location** | **Model** | **SICAPv2** | | **NFT** | | **NCT** | | **Average** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Acc | AUC | Acc | AUC | Acc | AUC | Acc | AUC |
| *Text-only* | | | | | | | | | | | |
| CoOp | Prompt | *input* level | CONCH | 59.80 | 82.74 | 90.52 | 96.81 | 88.97 | 98.58 | 79.76 | 92.71 |
| CoCoOp | Prompt | *input* level | CONCH | 63.24 | 85.02 | 90.92 | 96.48 | 93.05 | 99.49 | 82.40 | 93.66 |
| *Vision-only* | | | | | | | | | | | |
| CLIPath (RFC) | Adapter | *embed. level* | CONCH | 79.68 | 94.16 | 92.45 | 97.79 | 93.29 | 99.62 | 88.47 | 97.19 |
| LoRA$_{\text{Img}}$ | LoRA | *all layers* | CONCH | 71.16 | 91.32 | 79.40 | 93.39 | 95.11 | 99.50 | 81.89 | 94.74 |
| Linear Probe | Adapter | *embed. level* | CONCH | 79.45 | 93.80 | 92.71 | 97.91 | 95.47 | 99.27 | 89.21 | 96.99 |
| ProVP | Prompt | *all layers* | CONCH | 64.37 | 84.51 | 93.37 | 97.68 | 94.43 | 98.01 | 84.06 | 93.40 |
| *Multi-modal* | | | | | | | | | | | |
| MaPLe | Prompt | *shallow repr.* | CONCH | 66.87 | 90.23 | 92.19 | 97.24 | 93.53 | 99.60 | 84.20 | 95.69 |
| MMA | Adapter | *deep repr.* | CONCH | 43.70 | 71.71 | 49.97 | 65.46 | 23.23 | 68.02 | 38.97 | 68.40 |
| PromptSRC | Prompt | *all layers* | CONCH | 52.92 | 80.03 | 80.82 | 94.22 | 93.38 | 99.65 | 75.71 | 91.30 |
| MMRL | Mixed | *deep repr.* | CONCH | 70.88 | 87.06 | 94.34 | 98.02 | 95.50 | 99.17 | 86.91 | 94.75 |
| MMRL++ | Mixed | *deep repr.* | CONCH | 67.48 | 86.00 | 90.87 | 95.92 | 93.98 | 99.67 | 84.11 | 93.86 |

Table A8: Few-shot classification accuracy (%) of PEFT techniques on SICAPv2, NFT, and NCT using **CONCH**. Each dataset shows 1-shot (1S), 8-shot (8S), and 16-shot (16S) accuracy; **Avg** columns are per-shot means across datasets.

| **Method** | **Adaptation Type** | **Adaptation Location** | **SICAPv2** | | | **NFT** | | | **NCT** | | | **Avg** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1S | 8S | 16S | 1S | 8S | 16S | 1S | 8S | 16S | 1S | 8S | 16S |
| *Adapted modality: Text* | | | | | | | | | | | | | | |
| CoOp | Prompt | Input-level | 19.54 | 27.62 | 26.47 | 62.91 | 65.54 | 66.12 | 49.04 | 67.95 | 73.92 | 43.83 | 53.70 | 55.50 |
| CoCoOp | Prompt | Input-level | 27.41 | 35.42 | 32.96 | 66.26 | 62.65 | 65.36 | 32.54 | 52.99 | 61.17 | 42.07 | 50.35 | 53.16 |
| *Adapted modality: Image* | | | | | | | | | | | | | | |
| Linear Probe | Adapter | Embed./head-level | 39.68 | 61.77 | 64.70 | 69.61 | 72.26 | 71.19 | 76.34 | 90.97 | 93.72 | 61.88 | 75.00 | 76.54 |
| CLIPath (RFC) | Adapter | Embed./head-level | 35.34 | 61.92 | 70.14 | 62.06 | 70.20 | 67.70 | 52.11 | 88.28 | 90.57 | 49.84 | 73.47 | 76.14 |
| ProVP | Prompt | All layers | 31.97 | 41.86 | 40.07 | 75.71 | 62.64 | 65.10 | 38.78 | 49.60 | 50.21 | 48.82 | 51.37 | 51.79 |
| *Adapted modality: Multi-modal* | | | | | | | | | | | | | | |
| PromptSRC | Prompt | All layers | 33.11 | 41.27 | 47.59 | 42.83 | 55.46 | 61.37 | 49.05 | 61.92 | 70.36 | 41.66 | 52.88 | 59.77 |
| MaPLe | Prompt | Shallow repr | 28.28 | 34.13 | 31.79 | 31.04 | 41.85 | 52.67 | 18.96 | 44.52 | 54.46 | 26.09 | 40.17 | 46.31 |
| MMA | Adapter | Deep repr | 25.74 | 33.69 | 41.12 | 36.55 | 44.80 | 55.03 | 28.64 | 36.47 | 47.82 | 30.31 | 38.32 | 47.99 |
| MMRL | Adapter | Deep (late sem.) | 31.05 | 35.25 | 42.03 | 43.67 | 54.88 | 61.91 | 57.60 | 76.78 | 84.82 | 44.11 | 55.64 | 62.92 |
| MMRL++ | Adapter | Deep repr | 29.03 | 31.26 | 33.87 | 41.77 | 51.32 | 58.84 | 52.64 | 57.07 | 63.52 | 41.15 | 46.55 | 52.08 |

Table A9: Classification accuracy and AUCROC of **DFN** adapted using PEFT methods across SICAPv2, NFT, and NCT. Methods are stratified into text-only, image-only, and multimodal groups.

| **Method** | **Type** | **Location** | **Model** | **SICAPv2** | | **NFT** | | **NCT** | | **Avg** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Acc | AUC | Acc | AUC | Acc | AUC | Acc | AUC |
| *Text-only* | | | | | | | | | | | |
| CoOp | Prompt | *input* level | DFN | 68.57 | 88.64 | 90.26 | 96.43 | 92.92 | 99.50 | 83.92 | 94.86 |
| CoCoOp | Prompt | *input* level | DFN | 64.80 | 75.80 | 90.46 | 96.71 | 92.26 | 99.37 | 82.51 | 90.63 |
| *Vision-only* | | | | | | | | | | | |
| CLIPath (RFC) | Adapter | *embed. level* | DFN | 69.08 | 88.16 | 91.89 | 97.49 | 92.33 | 99.54 | 84.43 | 95.06 |
| LoRA$_{\text{Img}}$ | LoRA | *all layers* | DFN | 65.60 | 86.36 | 88.58 | 95.54 | 89.89 | 99.35 | 81.36 | 93.75 |
| Linear Probe | Adapter | *embed. level* | DFN | 67.58 | 87.51 | 91.59 | 97.12 | 87.70 | 98.58 | 82.29 | 94.40 |
| ProVP | Prompt | *all layers* | DFN | 74.36 | 92.00 | 93.52 | 97.33 | 92.68 | 99.52 | 86.85 | 96.28 |
| *Multi-modal* | | | | | | | | | | | |
| MaPLe | Prompt | *shallow repr.* | DFN | 75.58 | 93.14 | 93.52 | 97.79 | 92.30 | 97.88 | 87.13 | 96.27 |
| MMA | Adapter | *deep repr.* | DFN | 69.50 | 90.26 | 91.89 | 97.62 | 51.61 | 94.79 | 71.00 | 94.22 |
| PromptSRC | Prompt | *all layers* | DFN | 72.05 | 91.73 | 82.86 | 95.27 | 91.42 | 98.95 | 82.11 | 95.32 |
| MMRL | Mixed | *deep repr.* | DFN | 76.01 | 92.70 | 94.80 | 98.67 | 96.11 | 99.51 | 88.97 | 96.96 |
| MMRL++ | Mixed | *deep repr.* | DFN | 76.86 | 92.96 | 95.16 | 98.46 | 95.72 | 99.40 | 89.25 | 96.94 |

Table A10: Few-shot classification accuracy (%) of PEFT techniques on SICAPv2, NFT, and NCT using **DFN**. Each dataset shows 1-shot (1S), 8-shot (8S), and 16-shot (16S) accuracy; **Avg** columns are per-shot means across datasets.

| **Method** | **Adaptation Type** | **Adaptation Location** | **SICAPv2** | | | **NFT** | | | **NCT** | | | **Avg** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1S | 8S | 16S | 1S | 8S | 16S | 1S | 8S | 16S | 1S | 8S | 16S |
| *Adapted modality: Text* | | | | | | | | | | | | | | |
| CoOp | Prompt | Input-level | 27.95 | 36.00 | 45.00 | 56.57 | 60.65 | 61.02 | 53.03 | 72.66 | 77.44 | 45.85 | 56.44 | 61.15 |
| CoCoOp | Prompt | Input-level | 28.84 | 24.05 | 36.04 | 70.41 | 62.23 | 62.22 | 50.15 | 68.65 | 72.61 | 49.80 | 51.64 | 56.96 |
| *Adapted modality: Image* | | | | | | | | | | | | | | |
| Linear Probe | Adapter | Embed./head-level | 24.11 | 34.78 | 45.99 | 49.50 | 59.97 | 65.58 | 43.11 | 65.17 | 74.74 | 38.91 | 53.31 | 62.10 |
| CLIPath (RFC) | Adapter | Embed./head-level | 21.85 | 35.60 | 39.92 | 49.45 | 62.96 | 63.35 | 26.40 | 42.63 | 63.72 | 32.57 | 47.06 | 55.66 |
| ProVP | Prompt | All layers | 27.43 | 29.89 | 36.58 | 49.26 | 62.62 | 63.54 | 57.21 | 79.97 | 87.29 | 44.63 | 57.49 | 62.47 |
| *Adapted modality: Multi-modal* | | | | | | | | | | | | | | |
| PromptSRC | Prompt | All layers | 35.88 | 44.62 | 51.73 | 46.27 | 57.16 | 63.04 | 54.32 | 68.28 | 74.95 | 45.49 | 56.69 | 63.24 |
| MaPLe | Prompt | Shallow repr | 13.38 | 21.60 | 23.58 | 27.44 | 36.79 | 48.61 | 48.02 | 58.54 | 69.27 | 29.61 | 38.98 | 47.15 |
| MMA | Adapter | Deep repr | 38.80 | 38.80 | 37.80 | 25.46 | 34.64 | 42.73 | 30.53 | 34.23 | 39.43 | 31.60 | 35.89 | 39.99 |
| MMRL | Adapter | Deep (late sem.) | 30.92 | 43.92 | 50.38 | 44.33 | 53.47 | 59.62 | 58.24 | 82.22 | 87.47 | 44.50 | 59.87 | 65.82 |
| MMRL++ | Adapter | Deep repr | 26.46 | 45.95 | 51.57 | 47.75 | 61.95 | 62.95 | 60.96 | 82.38 | 87.10 | 45.06 | 63.43 | 67.21 |

Table A11: Classification Accuracy and AUCROC of **QuiltNet** + PEFT techniques after training on 100% of SICAPv2, NFT, and NCT datasets, with the adaptation location. An Avg column reports the mean Acc/AUC over the three datasets.

| **Method** | **Type** | **Location** | **Model** | **SICAPv2** | | **NFT** | | **NCT** | | **Avg** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Acc | AUC | Acc | AUC | Acc | AUC | Acc | AUC |
| *Text-only* | | | | | | | | | | | |
| CoOp | Prompt | *input* level | QuiltNet | 74.79 | 92.23 | 91.74 | 96.79 | 93.15 | 98.25 | 86.56 | 95.76 |
| CoCoOp | Prompt | *input* level | QuiltNet | 74.93 | 92.31 | 90.62 | 95.83 | 90.96 | 99.28 | 85.50 | 95.81 |
| *Vision-only* | | | | | | | | | | | |
| CLIPath (RFC) | Adapter | *embed. level* | QuiltNet | 72.62 | 90.91 | 92.29 | 97.53 | 71.21 | 95.48 | 78.71 | 94.64 |
| LoRA$_{\text{Img}}$ | LoRA | *all layers* | QuiltNet | 71.35 | 90.19 | 83.27 | 92.28 | 93.62 | 99.69 | 82.75 | 94.05 |
| Linear Probe | Adapter | *embed. level* | QuiltNet | 68.19 | 90.92 | 86.28 | 95.98 | 93.05 | 99.44 | 82.51 | 95.45 |
| ProVP | Prompt | *all layers* | QuiltNet | 74.00 | 91.42 | 94.79 | 97.53 | 92.47 | 99.38 | 87.09 | 96.11 |
| *Multi-modal* | | | | | | | | | | | |
| MaPLe | Prompt | *shallow repr.* | QuiltNet | 78.27 | 94.14 | 94.59 | 98.11 | 93.70 | 99.76 | 88.85 | 97.34 |
| MMA | Adapter | *deep repr.* | QuiltNet | 72.43 | 90.15 | 89.49 | 97.06 | 60.77 | 94.63 | 74.23 | 93.95 |
| PromptSRC | Prompt | *all layers* | QuiltNet | 75.49 | 93.05 | 76.74 | 93.20 | 94.05 | 99.45 | 82.09 | 95.23 |
| MMRL | Mixed | *deep repr.* | QuiltNet | 77.95 | 90.15 | 95.26 | 98.80 | 94.80 | 99.76 | 89.34 | 96.24 |
| MMRL++ | Mixed | *deep repr.* | QuiltNet | 77.76 | 89.81 | 95.87 | 98.97 | 95.54 | 99.79 | 89.72 | 96.19 |

Table A12: Few-shot classification accuracy (%) of PEFT techniques on SICAPv2, NFT, and NCT using **QuiltNet**. Each dataset shows 1-shot (1S), 8-shot (8S), and 16-shot (16S) accuracy; **Avg** columns are per-shot means across datasets.

| **Method** | **Adaptation Type** | **Adaptation Location** | **SICAPv2** | | | **NFT** | | | **NCT** | | | **Avg** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1S | 8S | 16S | 1S | 8S | 16S | 1S | 8S | 16S | 1S | 8S | 16S |
| *Adapted modality: Text* | | | | | | | | | | | | | | |
| CoOp | Prompt | Input-level | 41.67 | 57.01 | 58.00 | 59.60 | 64.13 | 63.76 | 67.49 | 81.97 | 85.35 | 56.25 | 67.70 | 69.04 |
| CoCoOp | Prompt | Input-level | 30.99 | 54.54 | 52.94 | 64.06 | 63.98 | 65.00 | 60.59 | 72.25 | 81.46 | 51.88 | 63.59 | 66.47 |
| *Adapted modality: Image* | | | | | | | | | | | | | | |
| Linear Probe | Adapter | Embed./head-level | 29.26 | 52.32 | 47.06 | 44.74 | 66.89 | 68.01 | 40.66 | 81.43 | 84.26 | 38.22 | 66.88 | 66.44 |
| CLIPath (RFC) | Adapter | Embed./head-level | 35.74 | 43.76 | 41.30 | 46.46 | 63.32 | 65.12 | 28.71 | 36.87 | 70.99 | 36.97 | 47.98 | 59.14 |
| ProVP | Prompt | All layers | 38.00 | 47.73 | 55.20 | 43.48 | 64.34 | 65.07 | 64.57 | 83.70 | 91.88 | 48.68 | 65.26 | 70.72 |
| *Adapted modality: Multi-modal* | | | | | | | | | | | | | | |
| PromptSRC | Prompt | All layers | 39.82 | 47.16 | 53.48 | 48.54 | 57.29 | 62.73 | 55.12 | 70.65 | 76.31 | 47.83 | 58.37 | 64.17 |
| MaPLe | Prompt | Shallow repr | 41.95 | 30.79 | 49.08 | 44.72 | 55.86 | 62.98 | 49.95 | 75.56 | 78.70 | 45.54 | 54.07 | 63.59 |
| MMA | Adapter | Deep repr | 23.80 | 23.87 | 44.53 | 32.26 | 44.37 | 55.01 | 32.90 | 52.63 | 76.43 | 29.65 | 40.29 | 58.66 |
| MMRL | Adapter | Deep (late sem.) | 43.89 | 55.42 | 61.78 | 52.77 | 63.92 | 70.46 | 73.68 | 89.42 | 91.64 | 56.78 | 69.59 | 74.63 |
| MMRL++ | Adapter | Deep repr | 44.85 | 61.26 | 58.13 | 50.44 | 60.37 | 67.21 | 72.10 | 88.06 | 90.49 | 55.80 | 69.90 | 71.94 |

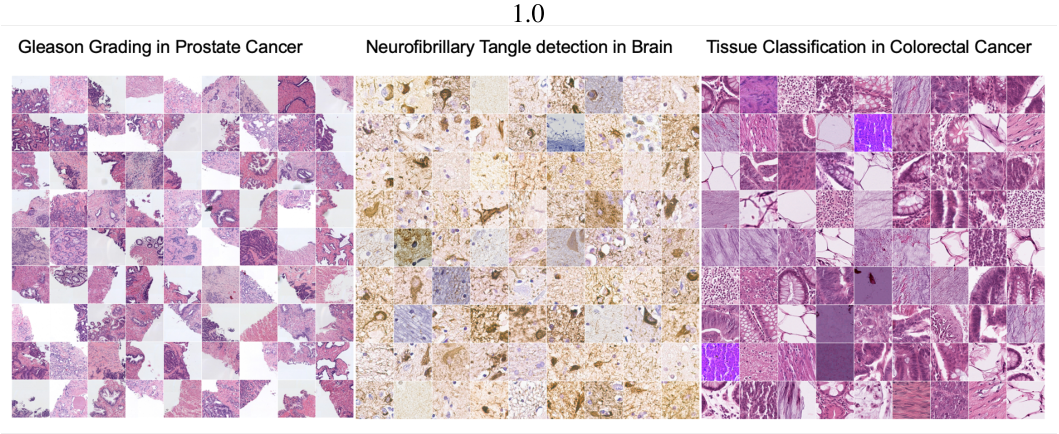
1.0
Gleason Grading in Prostate Cancer
Neurofibrillary Tangle detection in Brain
Tissue Classification in Colorectal Cancer

Figure A4: Randomly sampled images from three representative tasks and datasets.

the batch loss is,

$$\begin{aligned}\mathcal{L} = \tfrac{1}{2}\Bigg[\frac{1}{B}\sum_{i=1}^{B}\left(-\log\frac{\exp\big(\langle f_\theta(x_i),\, g_\phi(t_i)\rangle/\tau\big)}{\sum_{j=1}^{B}\exp\big(\langle f_\theta(x_i),\, g_\phi(t_j)\rangle/\tau\big)}\right) \\ + \frac{1}{B}\sum_{i=1}^{B}\left(-\log\frac{\exp\big(\langle g_\phi(t_i),\, f_\theta(x_i)\rangle/\tau\big)}{\sum_{j=1}^{B}\exp\big(\langle g_\phi(t_i),\, f_\theta(x_j)\rangle/\tau\big)}\right)\Bigg].\end{aligned} \tag{1}$$

Zero-shot classification forms class prototypes $z_c = g_\phi(t_c)$ from prompts $t_c$ and predicts $\hat{y} = \arg\max_c \langle f_\theta(x),\, z_c\rangle$.

**From web-scale pretraining to pathology.** Generic CLIP models learn from noisy web corpora. The domain gap to histopathology stems from ultra-high resolution, stain and scanner variability, texture-dominant cues, and specialized terminology Lai et al. [2024]. **CLIP** (and stronger OpenCLIP/DFN variants) provides the open-domain baseline; Domain-adapted VLMs aim to reduce this gap via curated image–text corpora: **QuiltNet** leverages the large curated Quilt-1M pathology corpus with template captions Ikezogwo et al. [2023]; **CONCH** mines pathology-specific captions for contrastive pretraining Lu et al. [2024]; and **PLIP/BioMedCLIP** are broader biomedical-tilted VLMs Huang et al. [2023], Zhang et al. [2023].

**Vision-only pathology foundational models :** Although there have been multiple foundational vision-only pathology models such as UNI [**?**] and Virchow [**?**]. They are beyond the scope for this work as we focus only on VLMs.

### C.2 Parameter-Efficient Fine-Tuning (PEFT)

Having already mentioned the PEFT strategies by their modalities (Axis A) in 2, we now classify PEFT strategies by their parameterization strategy: (*i*) prompt learning, (*ii*) adapter modules, (*iii*) low-rank updates (LoRA), and (*iv*) mixed methods. Each strategy modifies a frozen VLM differently, trading off efficiency, expressivity, and alignment.

### C.2.1 Prompt-based PEFT

Prompt-based methods learn additional tokens that steer the frozen text or vision encoder. The prototypical formulation comes from CoOp. Let $\mathbf{p} \in \mathbb{R}^{m\times d}$ denote $m$ learnable context vectors prepended to the textual embedding of class $y$. The logit for class $c$ is:

$$z_c(x) = \langle f_\theta(x),\, g_\phi([\mathbf{p}; \text{``}\texttt{class } c\text{''}])\rangle, \tag{2}$$

and the cross-entropy loss is:

$$\mathcal{L}_{\text{CoOp}}(x, y) = -\log\frac{\exp(z_y(x)/\tau)}{\sum_c \exp(z_c(x)/\tau)}. \tag{3}$$

Table A13: Taxonomy placement of PEFT methods arranged by **Axis A (What)** and ordered within each block by **Axis B (How)** and **Axis C (Where)**.

| Method | Axis A: Modality (What) | Axis B: Parameterization (How) | Axis C: Locus (Where) |
|---|---|---|---|
| *Text-only* | | | |
| CoOp | Text-only | Prompting | Input-level |
| CoCoOp | Text-only | Prompting | Input-level |
| *Image-only* | | | |
| ProVP | Image-only | Prompting | All layers |
| CLIPath (RFC) | Image-only | Adapters | Embed. / head-level |
| Linear Probe | Image-only | Adapters | Embed. / head-level |
| LoRA$_{\text{Img}}$ | Image-only | LoRA | Deep repr |
| *Multimodal* | | | |
| MaPLe | Multimodal | Prompting | Shallow repr |
| PromptSRC | Multimodal | Prompting | All layers |
| MMA | Multimodal | Adapters | Deep repr |
| MMRL | Multimodal | Mixed | Deep repr |
| MMRL++ | Multimodal | Mixed | Deep repr |

CoCoOp generalizes this by generating $\mathbf{p}(x)$ conditioned on image features, improving base-to-novel generalization. ProVP extends prompting into the vision encoder by inserting learnable tokens progressively across transformer blocks. PromptSRC introduces self-regularization to stabilize learned prompts under scarce labels for both text and vision encoders. MaPLe couples image and text prompts hierarchically to improve multimodal alignment.

### C.2.2 Adapter-based PEFT

Adapter methods insert light trainable layers into frozen encoders. A general formulation is given by CLIPath (RFC) [Lai et al., 2024], which fuses frozen CLIP features $f_\theta(x)$ with residual adapters $r_\psi(x)$:

$$f^* = \alpha\, L(f) + (1-\alpha)\, f, \tag{4}$$

We then replace $f$ with $f^*$ in the CLIP contrastive objective (Eq. 1) and optimize the trainable layers using the symmetric InfoNCE loss. This residual fusion connection (RFC) stabilizes learning in pathology with limited data.

CLIP-Adapter [Gao et al., 2024] implements a shallow MLP on visual features, while Tip-Adapter [Zhang et al., 2022] builds from cached training features for training-free adaptation. MMA [Yang et al., 2024] extends adapters to both image and text encoders, jointly fine-tuning lightweight modules across modalities.

### C.2.3 Low-Rank Update (LoRA)

LoRA [**?**] injects trainable low-rank decompositions into weight matrices. Given a frozen weight $W \in \mathbb{R}^{d\times k}$, LoRA reparameterizes it as:

$$W' \;=\; W \;+\; \Delta W, \qquad \Delta W \;=\; \frac{\alpha_{\text{lora}}}{r}\,AB, \quad A\in\mathbb{R}^{d\times r},\; B\in\mathbb{R}^{r\times k},\; r \ll \min(d,k), \tag{5}$$

while $r \ll \min(d,k)$. Only $A$ and $B$ are trainable; $\alpha_{\text{lora}}$ is a scaling hyperparameter. This significantly reduces parameters, and has been applied to both text and vision transformer blocks in VLMs. Variants include LoRA-Text, LoRA-Image, and multimodal LoRA, depending on placement.

### C.2.4 Mixed PEFT

Mixed methods like Representation-learning methods add regularizers that directly constrain multimodal embeddings, using a combination of multiple strategies like prompt-based and LoRA. MMRL introduces multimodal robust learning by aligning image and text adapters with consistency and distributional regularizers. Let $f_\theta(x)$ and $g_\phi(t)$ be frozen features, and $\tilde{f}_\psi(x), \tilde{g}_\psi(t)$ their adapted representations. MMRL minimizes:

$$\begin{aligned}\mathcal{L}_{\text{MMRL}} \;=\; & \mathcal{L}_{\text{cls}}\big(\tilde{f}_\psi(x),\, \tilde{g}_\psi(t);\, y\big) \\ & +\; \lambda_{\text{img}}\, \big\|\tilde{f}_\psi(x) \;-\; f_\theta(x)\big\|_2^2 \;+\; \lambda_{\text{txt}}\, \big\|\tilde{g}_\psi(t) \;-\; g_\phi(t)\big\|_2^2 \\ & +\; \lambda_{\text{align}}\, D\big(\tilde{f}_\psi(x),\, \tilde{g}_\psi(t)\big),\end{aligned} \tag{6}$$

where $\mathcal{L}_{\text{cls}}$ is the classification cross-entropy and $D(\cdot,\cdot)$ enforces multimodal agreement (e.g., cosine distance, MMD, or InfoNCE-style alignment).

MMRL++ simplifies these regularizers with fewer parameters and faster convergence, while preserving cross-modal robustness.

### C.3 Related Work: Benchmarks of PEFT+VLMs in Pathology

Recent surveys and benchmarks have begun to evaluate parameter-efficient finetuning (PEFT) in vision-language models (VLMs), but none provide a systematic, pathology-focused treatment. [Mai et al., 2025] unify PEFT strategies such as LoRA, adapters, and prompts across vision backbones, yet only include a single small pathology dataset (Camelyon) and one VLM configuration, without analyzing patch-based pathology, few-shot regimes, or alignment/localization under PEFT choices. [Lee et al., 2025] benchmark pathology FMs across 14–20 datasets and note that PEFT can be effective, but their scope is finetuning strategies for pathology-specific FMs, not systematic taxonomies of PEFT across CLIP-like VLMs or neuropathology tasks> They also do not provide evaluations for different types of PEFT methods. Related surveys catalog computational pathology (CPath) FMs and applications but similarly lack PEFT-structured comparisons or patch-level interpretability analyses Ochi et al. [2025], Chanda et al. [2024]. In contrast, our work is the first benchmark of PEFT+VLMs specifically for pathology, spanning cancer (SICAPv2, NCT) and neurodegeneration (NFT), across both open-domain and pathology-pretrained VLMs. We contribute a modality-aware taxonomy and a multi-dataset benchmark under both full-data and few-shot settings, explicitly linking accuracy vs. AUCROC trade-offs to efficiency (compute/storage), and uniquely incorporating alignment and localization insights for task alignment.

### C.4 Broader Impact

Parameter-efficient fine-tuning (PEFT) lowers the cost, memory, and expertise required to adapt vision–language models for histopathology. Our benchmark suggests that lightweight adapters/prompts/low-rank updates allow generalist models to approach—and sometimes exceed—pathology-pretrained models under full-data settings, while using a small fraction of trainable parameters and VRAM. This can broaden access for resource-constrained labs and modestly reduce environmental footprint. Our re-annotated NFT localization set also encourages evaluation beyond topline accuracy by testing whether models attend to pathologically meaningful regions. Our dataset is also already de-identified.

Risks remain: inconsistent localization on complex cases, a residual few-shot gap favoring pathology-pretrained models, domain shift across scanners/stains/sites, and annotation variability. We recommend (i) conservative, domain-bounded claims and research-only use; (ii) pairing accuracy with localization/uncertainty metrics; (iii) We report hardware and compute footprints to aid replication and impact estimation.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract/intro state the key claims—a pathology-focused PEFT taxonomy/benchmark across VLMs+PFMs, and that PEFT can let generalist VLMs rival or exceed PFMs under full-data, while PFMs keep a few-shot edge; these are borne out in the results. See Intro Sec 1 and Sec 3 Results.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper notes inconsistent localization in complex multi-NFT tiles and that pathology VLMs retain an advantage in extreme few-shot; it recommends alignment/localization evaluation beyond topline accuracy. See Conclusion, Sec 4.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] e

   Justification: Theoretical or formal proofs are not a contribution of this work.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification:We specify models, datasets, data regimes (full vs. 1/8/16-shot), training recipes (epochs, early stopping, optimizers, etc ), and PEFT-specific hyperparameters; few-shot runs are repeated over three seeds. B.2 Further, we will make our code repository and dataset avaialble for public use and reproducibility.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: At submission, no code/data URL is included. The paper relies on public images and introduces NFT localization annotations described in the text; code and novel annotation releases are planned post-review. Code will be made available at a GitHub repository. The author can not share the dataset and code at present, without violating confidentiality or annonymity. The dataset currently includes the annotators' personal information, code also includes substantial information about the authors. This information will be scrubbed from the dataset and **will be provided for public use.**

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Data preparation/splits, prompts, training schedules, and method-specific settings (e.g., token counts, ranks, adapter depths) are documented; evaluation metrics (Acc/AUC) are defined. See Appendix B.2 and Section 2

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We report averages over three seeds for few-shot but do not include error bars/confidence intervals for brevity sake; adding stdev/CIs is feasible in the camera-ready.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report GPU types (T4 16 GB; L40S 48 GB) and efficiency metrics (GFLOPs, trainable params, peak VRAM by method; 2), but not wall-clock time or total compute for each run. We can provide that in the camera-ready version, it was omitted for brevity sake as it did not affect the messaging of the paper. The full research project did not require more compute than mentioned, and we do report compute workers, relevant memory and storage.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Work uses de-identified pathology datasets and expert re-annotation; no personal data or interventions with human subjects; anonymity preserved for review.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss positive impacts (lowering compute/memory barriers, alignment-aware evaluation) and risks (domain shift, miscalibration, reduced reliability in complex cases, inconsistent localization) with mitigations; see Broader Impact Appendix section C.4 and Conclusion section4.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No high-risk generative models or scraped web corpora are released with this submission; any planned assets will follow controlled and de-identified release practices.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: No licence provided by [Vizcarra et al., 2023] - `https://github.com/Gutman-Lab/yolo-braak-stage`, [Ghandian et al., 2024] make their data available at `https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BIAD1165` under `https://creativecommons.org/publicdomain/zero/1.0/legalcode` CCO licence. We credit and cite them, as well as properly respect the terms of their licence. The authors of both the studies were reached out to, for the purposes of this study and duly informed about the usage of their datasets.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: The NFT annotation resource is introduced and documented (consensus/union labels, inter-rater variability), but no packaged release/docs accompany the submission; a dataset card will be provided upon release.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing or human-subject experiments were conducted; annotations were made by domain experts on de-identified images.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The study analyzes existing de-identified images with expert re-annotation only; no recruitment or intervention with human participants.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The research methods and results do not rely on LLMs for experiments; any editing assistance is unrelated to methodology.