

From Classification to Cross-Modal Understanding: Leveraging Vision-Language Models for Fine-Grained Renal Pathology

Zhenhao Guo ¹, Rachit Saluja ², Tianyu Shi ³, Tianyuan Yao ⁴, Quan Liu ⁴, Junchao Zhu ⁴, Haibo Wang ⁵, Daniel Reisenbüchler ⁶, Benjamin Liechty ⁷, David J. Pisapia ⁷, Kenji Ikemura ⁷, Steven Salvatoree ⁷, Surya Seshane ⁷, Mert R. Sabuncu ^{2,7}, Yihe Yang ^{7,8}, Ruining Deng ^{4,7}

¹ New York University, New York, NY 10012, USA

² Cornell Tech, New York, NY 10044, USA

³ Sichuan University, Chengdu, 610207, CN

⁴ Vanderbilt University, Nashville, TN, 37235, USA

⁵ Carnegie Mellon University, Pittsburgh, PA 15213, USA

⁶ University of Regensburg, Regensburg, Bavaria 93053, DE

⁷ Weill Cornell Medicine, New York, NY 10065, USA

⁸ Northwell Health, New Hyde Park, NY 11040, USA

Abstract

Fine-grained glomerular subtyping is essential for kidney biopsy interpretation, yet it presents a harder setting than the broad organ- or disease-level recognition tasks commonly used in vision-language model (VLM) evaluation. In this task, all categories arise within the same renal microanatomic unit, and clinically meaningful subtypes often differ by subtle, overlapping, and continuum-like morphological cues. Clinically valuable labels are also scarce, making fully supervised image-only paradigms difficult to translate to practice. Although VLMs are promising for data-efficient pathology learning, it remains unclear how their cross-modal representations should be adapted and evaluated for such within-structure discrimination under realistic few-shot constraints. To our knowledge, this study establishes the first fine-grained renal pathology benchmark for few-shot VLM adaptation and systematically examines how supervision level, model design, and adaptation strategy shape both classification performance and multimodal representation structure. Beyond accuracy, area under the receiver operating characteristic curve (AUC), F1 score, image-text alignment, positive-negative similarity separation, embedding geometry, and confidence calibration quantified by expected calibration error (ECE), we ask when performance gains reflect more reliable cross-modal understanding. Across renal pathology cohorts, additional supervision consistently improves discrimination, but reliable adaptation is not explained by alignment alone; stronger performance is associated with larger class separation, more coherent embedding organization, and better awareness of confidence-related failure modes under domain variation. Within this retrospective two-cohort evaluation, these findings move assessment beyond coarse classification and accuracy alone and provide comparative evidence to inform model selection and future translational validation. Our implementation has now been publicly released for reproducibility on GitHub at <https://github.com/ddrrnn123/glo-vlms>.

Keywords

Vision-Language Model, Fine-Grained Classification, Fine-tuning, Few-shot Learning, Digital Renal Pathology

Article informations

<https://doi.org/10.59275/j.melba.2026-3893>

Volume 2026, Received: 2025-11-14, Published yyyy-m2-d2

Corresponding author: rud4004@med.cornell.edu

©2026 Zhenhao Guo et al. License: CC-BY 4.0



1. Introduction

Renal histopathology is central to diagnosis, subtyping, and prognosis in kidney disease Feng and Liu (2023); Sethi and Fervenza (2019); Schnuelle (2023); Wilson and Messias (2022). Fine-grained glomerular subtypes carry therapeutic and risk-stratification implications beyond coarse lesion detection Working Group of

the International IgA Nephropathy Network and the Renal Pathology Society et al. (2009); Rafieian-Kopaei et al. (2013); Chen et al. (2022); Lu et al. (2025); Deng et al. (2022); Yu et al. (2025). Yet high-quality labels are expensive, inter-expert agreement is limited, and cross-institution variation in staining and scanners induces domain shift Duenweg et al. (2023). In nephrectomy cohorts of renal cell carcinoma (RCC), surgical clamping of the renal artery, veins, or ureter produces ischemic artifacts such as glomerular basement membrane (GBM) wrinkling and capsular distention, so many glomeruli share similar ischemic-appearing changes. RCC patients also share common risk factors with diabetes, hypertension, and metabolic syndrome, and many glomeruli lie along a continuum from ischemia to segmental and then global sclerosis, often showing GBM wrinkling, multilayering of Bowman's capsule, and tuft-capsule adhesions in the same glomerulus. These surgical and clinical factors blur the boundaries between ischemic, segmentally sclerotic, and globally sclerotic glomeruli, making it difficult for algorithms to distinguish between these groups. In practice, small labeled cohorts are the norm, and clinicians often reason with a few confirmed exemplars as references Crowley et al. (2003). Recent advances in foundation VLMs have transformed computational pathology, offering generalizable representations that bridge visual and textual modalities Cui et al. (2022); Chanda et al. (2024); Bilal et al. (2025); Ochi et al. (2025). However, standardized evidence for their fine-grained adaptation remains limited, complicating principled selection across backbones and parameter-efficient fine-tuning strategies. Prior studies typically evaluate only a single task granularity Campanella et al. (2019), supervision regime Lu et al. (2021), or adaptation strategy at a time Yao et al. (2022), often focusing on coarse normal-vs-diseased discrimination Wang et al. (2025), training under full supervision Yu et al. (2025), or image-only classification with one backbone and one split Yao et al. (2022), and rarely characterizing how performance scales with sample size Taher et al. (2025); Hosseinzadeh Taher et al. (2021). At the same time, the degree to which visual and textual features are aligned is increasingly viewed as central to the robustness and clinical reliability of VLMs in pathology Chen et al. (2024); Li and Tang (2024); Hanna et al. (2025). We address these gaps by (i) establishing a framework for fine-grained glomerular subtyping, (ii) characterizing model performance in a clinically realistic few-shot regime, and (iii) analyzing multimodal image-text feature alignment and between-class separability. We evaluate both general-purpose and pathology-specialized backbones across prevailing architectures and multiple parameter-efficient adaptation strategies across few-shot settings using Monte Carlo resampling. Beyond measuring performance with classification metrics, we perform a cross-modal embedding analysis: we quantify image-text feature alignment and

between-class separability as a function of supervision level, task granularity, and adaptation strategy. We then relate these representational properties to model performance and ECE. Taken together, this reveals how supervision and adaptation jointly shape multimodal alignment and downstream behavior, providing practical guidance for model selection under realistic data constraints.

Our contributions are threefold:

1. To our knowledge, we establish the first VLM benchmark for fine-grained glomerular subtyping in renal pathology, targeting pathologically actionable distinctions beyond coarse disease classification.
2. We characterize model behavior under realistic data limitations by restricting supervision to a few labeled exemplars per class and analyzing how backbone, parameter-efficient adaptation, and sample size jointly affect performance in general-purpose and pathology-specialized foundation models.
3. We introduce embedding-space analyses that quantify multimodal image-text feature alignment and between-class separability, and we show that these metrics track downstream accuracy and confidence calibration, revealing trends that inform future cross-modal learning under few-shot constraints.

2. Related Works

2.1 Few-shot learning for medical image classification

Renal pathologists have established morphology-based systems for fine-grained glomerular classification. Recent work has further refined outcome assessment in focal segmental glomerulosclerosis through a new index Tervaert et al. (2010); Bajema et al. (2018); Chan et al. (2024). Together, these expert-defined schemas provide fine-grained and clinically meaningful labels that now serve as reliable ground truth for training AI models. However, pathology data are inherently scarce, costly to annotate, and highly heterogeneous across institutions and imaging protocols Madabhushi and Lee (2016); Shi et al. (2019); Li et al. (2025a). As a result, few-shot learning naturally emerges as a practical and clinically aligned approach that maintains robust performance with only a handful of examples per class. Prior studies indicate that such methods can substantially mitigate data scarcity and even approach fully supervised accuracy with very limited samples Pachetti and Colantonio (2024); Shakeri et al. (2022). Accordingly, few-shot learning provides a compelling direction for developing reliable and data-efficient pathology AI systems under low-annotation constraints. However, prior work rarely reports shot-scaling behavior or evaluates adaptation under explicit cross-center

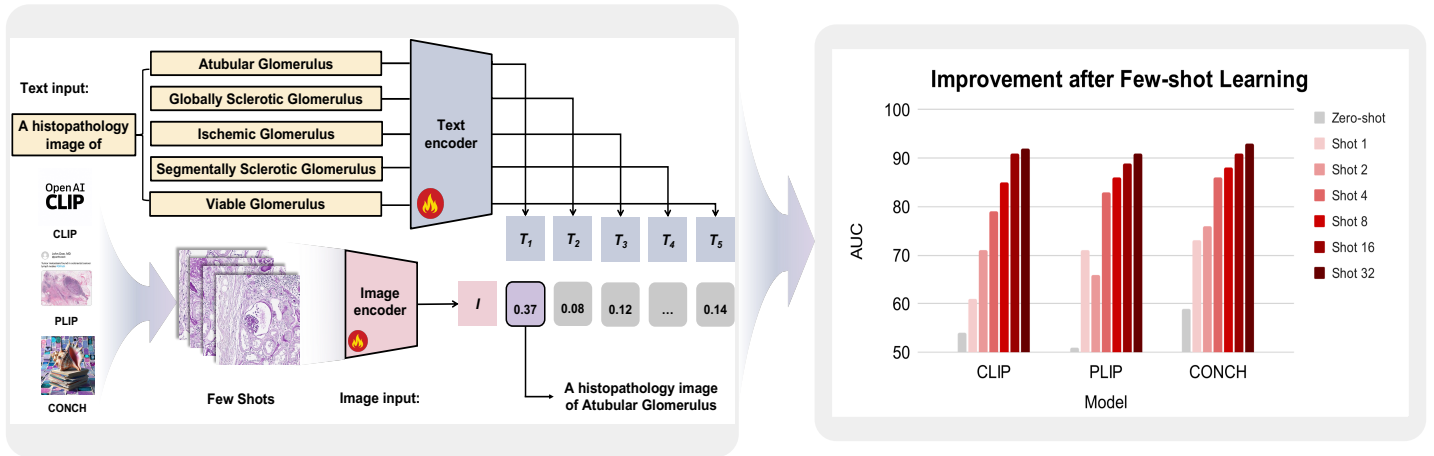


Figure 1: **Pipeline overview.** We train the image encoder and the text encoder in a few-shot setting, where only a small number of labels are provided for each glomerular subtype. For each shot, patches and their class descriptions are encoded, and the model is updated. As the number of shots increases, the performance improves.

domain shift, leaving unclear how shot budget affects robustness and which adaptation strategy remains reliable.

In our work, we treat fine-grained renal pathology classification as a clinically grounded few-shot medical image classification problem and disentangle the effects of three key factors: foundation model choice, fine-tuning strategy, and shot budget. Under controlled whole-slide image (WSI)-level splits and repeated runs, we systematically vary the backbone (general-purpose and pathology-specialized), the adaptation method (full vs parameter-efficient updates), and the number of labeled examples per class, and report shot-scaling behavior across these settings. This study is designed to provide practical insights into which combinations are most data-efficient and how additional annotations translate into measurable gains, informing model selection and annotation investment under realistic data constraints.

2.2 Fine-grained Glomerular Classification

In recent years, deep learning and VLMs have achieved substantive progress in fine-grained glomerular classification. Deep learning, grounded in morphological patterns, has produced end-to-end convolutional pipelines on WSIs that cover glomerulus detection, disease categorization, and key lesion identification Yao et al. (2022); Lu et al. (2021); Yu et al. (2025); Weis et al. (2022); Lei et al. (2024); Li et al. (2025b). At the same time, the feasibility of VLMs for pathology classification has been increasingly demonstrated. These VLM-based methods have shown strong performance Zheng et al. (2025); Guo et al. (2026). Taken together, these studies provide converging evidence that both deep learning and VLMs are feasible for fine-grained glomerular classification. However, most existing studies emphasize Top-1 performance, while practical deployment constraints and reliability measures are rarely reported in

a unified and systematic manner. In particular, confidence calibration (e.g., ECE) and decision-margin behavior (e.g., boundary margin) are seldom quantified, limiting evidence about confidence reliability under clinical conditions. Consequently, it remains unclear how these methods translate into dependable and deployable decision support in renal pathology workflows.

To address these limitations, we perform a systematic evaluation of few-shot fine-grained glomerular classification under clinically realistic data constraints. In our work, we compare general-purpose and pathology-specialized backbones as well as multiple adaptation strategies under matched shot budgets and cross-center validation, and we report not only discrimination metrics (Accuracy, AUC, and F1) but also confidence calibration. Moreover, we measure ECE and boundary margin to characterize whether predicted probabilities are calibrated and whether decisions are made with a sufficient separation between competing classes. This combined view provides a broader research assessment of classification performance and confidence behavior under limited supervision and cross-center variation.

2.3 Cross-modal Vision Language Models on Renal Pathology

In clinical practice, physicians engage in a multimodal and integrative diagnostic process, synthesizing information from diverse sources. This cross-modal verification ensures that diagnostic reasoning is both reliable and interpretable Roth et al. (2021); Cui et al. (2024); Jin et al. (2025). Analogously, to endow artificial intelligence systems with comparable credibility in medical contexts, it is essential to move beyond unimodal classification paradigms and promote modality alignment. Such alignment enables the model to ground its predictions in coherent morphological

and textual evidence, thereby enhancing transparency and interpretability Kline et al. (2022); Simon et al. (2025). Moreover, effective modality alignment contributes to improved empirical performance, as it facilitates the extraction of stable and domain-general representations within a unified semantic space Wang and Isola (2020); Xu et al. (2024); Wang and Carvalho (2026, 2023). Despite rapid progress in multimodal learning and pathology VLMs, most prior studies treat downstream metrics as primary endpoints. Consequently, feature-level analyses of cross-modal representations in fine-grained renal pathology remain limited, including (i) explicit alignment diagnostics, (ii) class-wise similarity gaps/margins, and (iii) joint visualization of image and text embeddings.

To address these limitations, we explicitly evaluate multimodal representation quality alongside downstream classification performance. In our work, we adapt several general and pathology-specialized VLM backbones to fine-grained glomerular subtyping in a clinically realistic few-shot regime and track how different adaptation strategies reshape the joint image-text embedding space. Beyond reporting Accuracy, AUC, and F1, we quantify paired image-text alignment, class-wise similarity separation, and embedding geometry through visualization and clustering statistics, and we examine how these indicators co-vary with model confidence calibration and cross-center generalization. This feature-level perspective provides a more interpretable basis for selecting VLM backbones and fine-tuning strategies under limited supervision and domain shift.

2.4 Pathology Benchmarking and Challenge

In digital pathology, community-driven grand challenges and benchmarks have been key drivers of methodological progress by providing standardized datasets and evaluation protocols that enable objective, like-for-like comparison and improve reproducibility. For example, CAMELYON16/17 established a standardized setting for WSI lymph-node metastasis detection and classification, enabling systematic benchmarking of WSI-level tumor recognition methods Ehteshami Bejnordi et al. (2017); Bandi et al. (2019). More recently, Glo-VLMs extended benchmark-style evaluation to fine-grained renal pathology classification under few-shot regimes, offering a structured view of current VLM capabilities and their limits Guo et al. (2026); Sohail et al. (2025). Importantly, Glo-VLMs also highlighted remaining evidence gaps and called for stronger feature-level diagnostic analyses, repeated runs to assess stability, and evaluation on external datasets to probe cross-center generalization. These observations align with recent surveys emphasizing the importance of feature analysis, few-shot learning, fine-grained recognition, and cross-center robustness in medical imaging Xiong et al. (2025); Hosseini et al. (2024); Nazir et al.

(2023); Yoon et al. (2024). However, existing evaluations typically study these dimensions in isolation across different datasets and experimental settings, lacking a unified and reproducible benchmarking protocol that jointly characterizes these key variables. This fragmentation makes method selection difficult, undermines the stability and comparability of reported gains, and ultimately weakens evidence for clinical usability.

To address this gap, we introduce a unified benchmark for fine-grained renal pathology classification: under controlled few-shot settings and cross-center external validation, we systematically compare mainstream VLMs and multiple adaptation strategies, and complement downstream metrics with feature-level diagnostic analyses to characterize the performance, representation and generalization relationship, providing a reproducible basis for method comparison and future translational evaluation.

3. Methods

3.1 Pipeline Overview

To overcome the limited generalizability of existing methods for fine-grained pathological image classification, our approach centers on distinguishing among subtypes of diseased glomeruli within a low-data setting. Specifically, we investigate the adaptation of VLMs to a five-class glomerular pathology classification task, relying on only a small number of labeled examples per category. As depicted in Fig. 1, the pipeline begins with pathological image patches paired with their corresponding class prompts (for example, “A histopathology image of Atubular Glomerulus”). Both the images and textual prompts are encoded using a pretrained VLM backbone, which is adapted to the task either through fine-tuning or the integration of lightweight modules. To reflect the low-data scenario, this adaptation is performed under few-shot conditions, where only a limited set of labeled exemplars per class is available to guide the alignment between modalities. This process yields high-dimensional embeddings for both images and prompts. Except for Classifier tuning, all methods use image-text cosine similarity for inference: we compute the cosine similarity between each image embedding and all class prompt embeddings, applying a softmax transformation to produce a probability distribution over the classes. The model then assigns each image to the class with the highest similarity score. For Classifier tuning, predictions are instead produced by a learned classification head on frozen image features.

3.2 Foundation VLM Backbones

To systematically assess the influence of model design, domain expertise, and modality alignment, we select three VLM backbones, as illustrated in Fig. 2 parts a and b, that

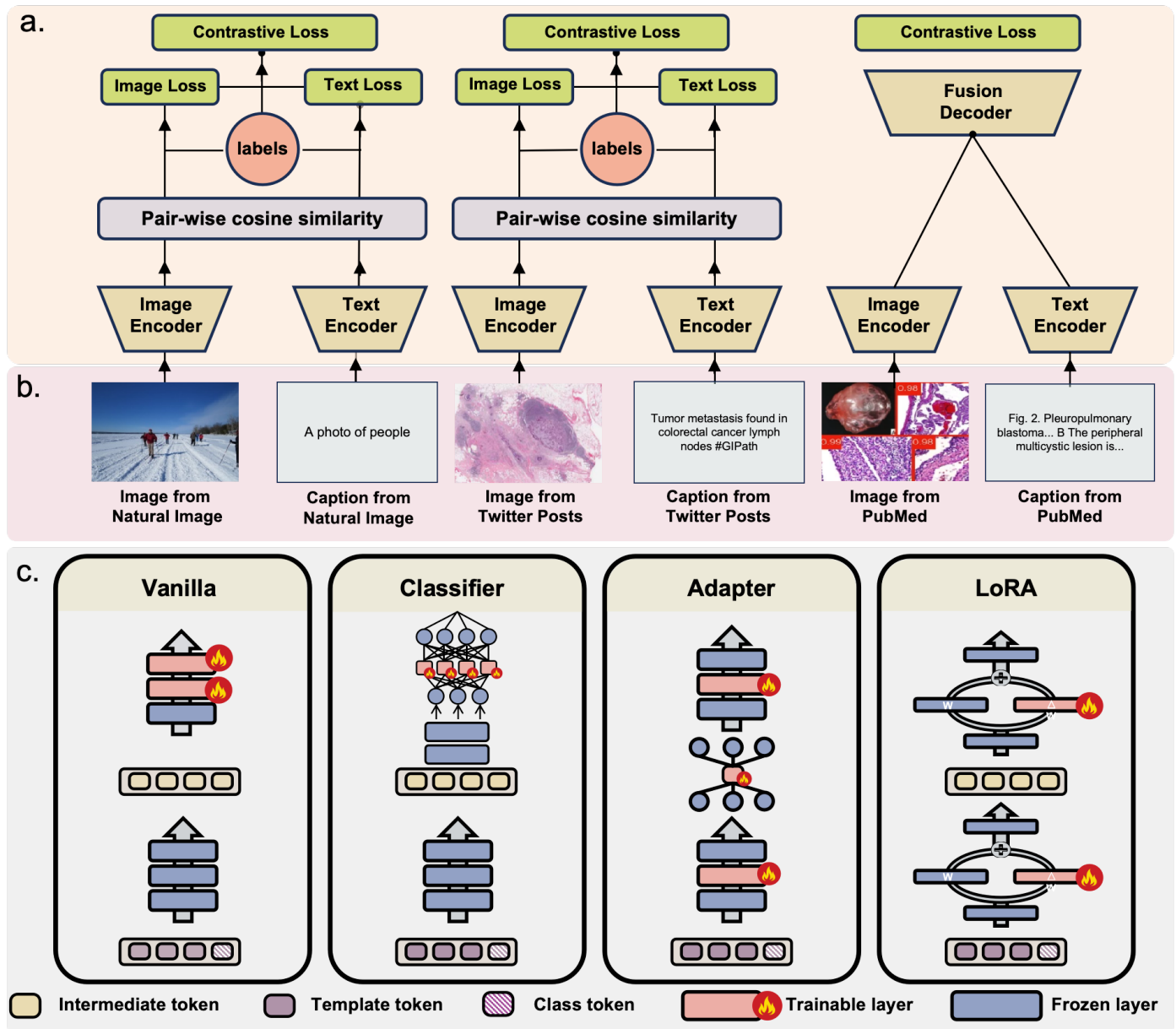


Figure 2: **(a) Model architectures.** We compare contrastive VLMs with separate image and text encoders trained via pairwise cosine similarity, and a CoCa-style architecture with an additional fusion/decoder branch. **(b) Domain-specific training data.** We illustrate three sources of paired image-text data: natural images with generic captions, social media posts from Twitter with informal expert commentary, and biomedical figures from PubMed with pathology-style captions. **(c) Fine-tuning strategies.** We compare (i) Vanilla fine-tuning of the top layers, (ii) LoRA low-rank updates, (iii) trainable adapters between frozen layers, and (iv) a classifier head on a frozen backbone. VLM adaptation uses only the image-text contrastive loss, whereas Classifier tuning uses cross-entropy; no captioning loss is used.

differ in scale, architecture, and the extent of pathology-specific training. These are: a general-purpose foundation model (CLIP), a pathology-adapted CLIP model (PLIP), and a pathology-native foundation model (CONCH).

- **CLIP:** A large-scale VLM Radford et al. (2021), trained on roughly 400M images with natural language captions using a symmetric contrastive learning objective. In this work, we adopt the openai/clip-vit-base-patch16 variant, which employs a ViT-B/16 image encoder and a

Transformer-based text encoder Dosovitskiy et al. (2020). Both encoders project their respective inputs into a shared embedding space, normalized for cosine similarity. CLIP enables zero-shot classification and retrieval by comparing the embeddings of images against those of natural language prompts for similarity-based matching.

- **PLIP:** A pathology-adapted VLM Huang et al. (2023), developed by retraining CLIP on over 200K pathology image-caption pairs collected primarily from Twitter posts.

The model retains the dual-encoder architecture of CLIP, with a ViT-B/16 image encoder and a Transformer-based text encoder, and projects both modalities into a shared embedding space. By aligning histopathology images with domain-specific textual descriptions, PLIP produces embeddings that are more sensitive to pathology-specific features and supports zero-shot classification and cross-modal retrieval in the medical imaging domain.

- **CONCH:** A pathology-native VLM Lu et al. (2024), developed on the CoCa architecture Yu et al. (2022) with a ViT-B/16 backbone. It is pretrained on 1.17M pathology image-caption pairs curated from PubMed and related pathology resources, combining contrastive alignment and caption generation objectives. By integrating structured attention pooling and a fusion decoder within the CoCa framework, CONCH learns fine-grained multimodal representations of high-resolution histology and supports zero-shot classification and cross-modal retrieval in the pathology domain. In our experiments, we do not use a captioning loss; we fine-tune CONCH using only image-text contrastive learning with fixed class prompts.

3.3 Fine-tuning Strategies

To adapt VLMs for fine-grained glomerular classification in few-shot settings, we examine four fine-tuning approaches, as illustrated in Fig. 2 part c, spanning full-parameter updates to highly parameter-efficient methods Gao et al. (2024); Srinivasan et al. (2024); Hu et al. (2022); Ioffe and Szegedy (2015). Vanilla fine-tuning modifies both image and text encoders, while Adapter and LoRA insert lightweight modules into the encoder to adjust internal representations with minimal disruption. Classifier tuning updates only the prediction head, keeping the backbone fixed. For VLM adaptation methods (Vanilla, Adapter, and LoRA), we compute class probabilities via cosine similarity between image and text embeddings followed by softmax. Since our task is classification with fixed prompts (not caption generation), we optimize only the image-text contrastive loss and do not use any captioning loss. For Vanilla, Adapter, and LoRA, we use the standard symmetric CLIP image-text contrastive objective Radford et al. (2021). Let $\mathbf{S} \in \mathbb{R}^{N \times N}$ denote the matrix of temperature-scaled cosine similarities for a batch of N paired image and text embeddings, and let $\mathbf{y}^{\text{pair}} = (1, \dots, N)$ denote the diagonal pairing targets. The final objective is

$$\mathcal{L}_{\text{ITC}} = \frac{1}{2} \left[\text{CE}(\mathbf{S}, \mathbf{y}^{\text{pair}}) + \text{CE}(\mathbf{S}^{\top}, \mathbf{y}^{\text{pair}}) \right]. \quad (1)$$

The normalized embeddings, similarity scores, and directional loss terms are detailed in Appendix A. Classifier

tuning uses standard multiclass cross-entropy; its complete definition is also provided in Appendix A.

- **Vanilla Fine-tuning Srinivasan et al. (2024):** This method unfreezes the top layers of the image and text encoders and jointly updates their parameters along with the task-specific head. By training these layers end-to-end, the model adjusts its high-level representations directly to the target task. However, this flexibility comes at the cost of higher computational demands.
- **Low-Rank Adaptation (LoRA) Hu et al. (2022):** This method injects additional trainable low-rank decomposition matrices into the attention projection layers of the top transformer blocks in both the vision and text encoders. During training, only low-rank decomposition matrices are updated while the original backbone weights remain frozen. The rank r controls the dimensionality of the adaptation subspace, and the scaling factor α regulates the contribution of ΔW to the final projection. This design reduces the number of trainable parameters while preserving the pretrained backbone.
- **Adapter Tuning Gao et al. (2024):** This approach preserves all pretrained encoder weights and introduces lightweight bottleneck adapter modules into the Transformer layers Pfeiffer et al. (2020). In our setup, we adopt *sequential bottleneck adapter*, which places a single adapter block only after the feed-forward network of each Transformer layer. Each adapter consists of a down-projection from the hidden dimension d to a smaller bottleneck dimension m , a non-linear activation, and an up-projection back to d . The transformed output is then added to the residual stream. During training, only the adapter parameters are updated while the backbone remains frozen, enabling parameter-efficient adaptation with minimal additional overhead.
- **Classifier Tuning Ioffe and Szegedy (2015):** This method freezes all backbone parameters and trains only a lightweight classification head on top of the extracted image features. We evaluate three variants: (i) a single linear projection from the feature dimension to the number of classes; (ii) a two-layer multilayer perceptron (MLP) with a hidden layer and non-linear activation Guo et al. (2025); and (iii) the same two-layer MLP augmented with batch normalization Ioffe and Szegedy (2015) applied after the hidden layer. In all cases, the classifier outputs class logits which are passed through a softmax for prediction. For all reported Classifier tuning results, the final classifier head maps the 512-dimensional frozen image feature to a 512-dimensional hidden layer, followed sequentially by batch normalization, a ReLU activation, and dropout, before a final linear layer pro-

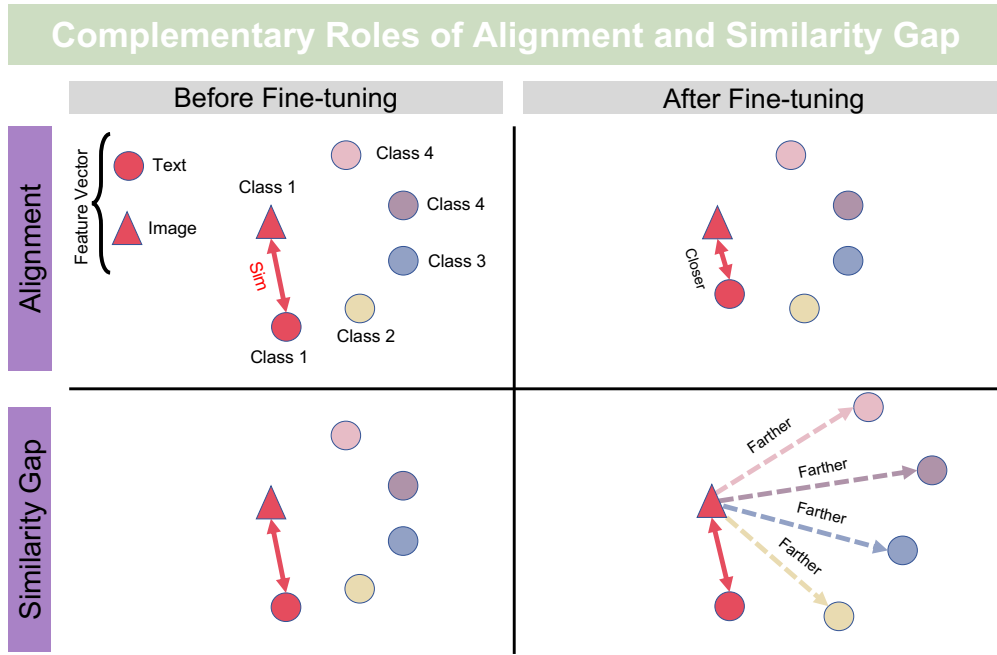


Figure 3: **Complementary roles of alignment and similarity gap.** It shows the change in image and text feature vectors before and after fine-tuning. Alignment measures whether fine-tuning pulls an image embedding closer to the embedding of its own positive text description. Similarity gap is defined as mean positive class-prompt similarity (Alignment) minus mean negative class-prompt similarity, and therefore increases when matched pairs become closer and/or unmatched class prompts become less similar. Together, these two effects capture how fine-tuning reduces within-class image-text distance while enlarging across-class distance.

duces the C class logits. In our experiments, variant (iii) performs best, and we use it for all reported results.

(502, 7.46%), Segmental glomerulosclerosis (271, 4.03%), and Viable glomeruli (3,223, 47.92%).

We additionally include two image-only linear-probe baselines, ImageNet-pretrained ResNet-18 He et al. (2016) and histopathology-specific self-supervised learning (SSL) ResNet-18 Ciga et al. (2022), to contextualize the benefits of cross-modal pretraining.

4. Experiments and Data

4.1 Data and Composition

We conducted experiments on two renal WSI cohorts, referred to as the Cornell dataset and the Vanderbilt dataset.

4.1.1 Cornell Dataset

The Cornell cohort retrospectively includes the non-tumoral kidney parenchyma of eight radical nephrectomy cases, yielding a total of 17 WSIs, processed at the Department of Pathology, Northwell Health, with specimens sourced from Long Island Jewish Medical Center and Northshore University Hospital (July 2017 to October 2019). Using the provided annotation masks, we extracted 6,726 glomerulus-centered patches and assigned each patch to one of five glomerular subtypes: Atubular glomeruli (230, 3.42%), Global glomerulosclerosis (2,500, 37.17%), Ischemic glomeruli

4.1.2 Vanderbilt Dataset

The Vanderbilt cohort comprises 141 renal WSIs from 126 patients with linked patient-level clinical information. Tissue specimens were processed as formalin-fixed, paraffin-embedded blocks, sectioned at $3\mu\text{m}$, and stained with Periodic Acid-Schiff. The cohort included 87 males and 39 females (mean age 63.4 ± 12.0 years; median 63.8 years) and was predominantly Caucasian (92.9%), followed by African American (6.3%) and Asian-Pacific (0.8%). Available clinical annotations included serum creatinine, blood urea nitrogen, creatinine clearance, hypertension, and diabetes; mean (median) serum creatinine, blood urea nitrogen, and creatinine clearance were 1.68 (1.02) mg/dL, 22.82 (18.00) mg/dL, and 66.42 (72.74) mL/min, respectively. From the WSIs, we extracted 13,922 patches and assigned each patch to one of five categories: Disappearing glomeruli (434, 3.12%), Non-glomerular (4,865, 34.94%), Normal glomeruli (3,609, 25.92%), Obsolescent glomeruli (4,289, 30.81%), and Solidified glomeruli (725, 5.21%).

ROC Comparison of Backbones and Fine-tuning Strategies Across Shots

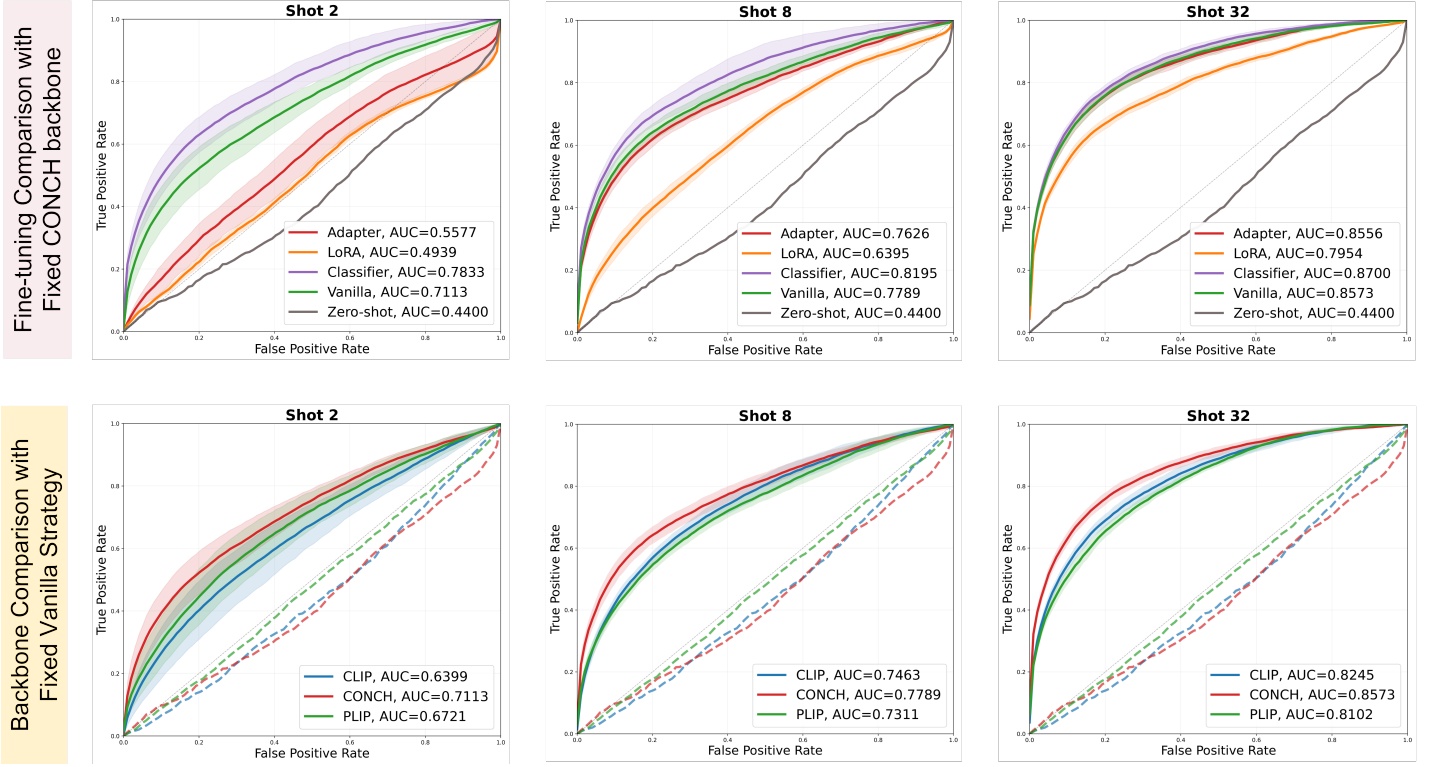


Figure 4: **Receiver operating characteristic (ROC) curves across models and fine-tuning methods on Cornell dataset.** We compare two axes, model comparison (fixed strategy) and fine-tuning method comparison (fixed backbone), by plotting true-positive rate (TPR) versus false-positive rate (FPR) across increasing shot counts.

4.2 Data Preprocessing

To standardize both the physical sampling resolution and the model input dimensions across cohorts, image patches were generated at a common target resolution of approximately $0.50 \mu\text{m}/\text{pixel}$. The Cornell WSIs were acquired at $20\times$ magnification with a native pixel spacing of approximately $0.50 \mu\text{m}/\text{pixel}$, whereas the Vanderbilt WSIs were acquired at $40\times$ magnification with a native pixel spacing of approximately $0.25 \mu\text{m}/\text{pixel}$, as recorded in the WSI metadata. During patch extraction, Vanderbilt image regions were downsampled to approximately $0.50 \mu\text{m}/\text{pixel}$ to match the Cornell cohort. Individual glomeruli were localized using the provided annotation masks, and an annotation-derived bounding box was computed for each glomerulus. The center of this bounding box was then used as the center of a fixed $512 \mu\text{m} \times 512 \mu\text{m}$ square field of view. Sampling this field at approximately $0.50 \mu\text{m}/\text{pixel}$ produced a 1024×1024 -pixel patch for downstream training. This procedure ensured consistent physical sampling resolution and field of view across both cohorts.

For the textual component, we generate class-specific prompts that serve as anchors for contrastive learning. Each of the five glomerular subtypes is represented by a stan-

dardized phrase in the form: “A histopathology image of {label}.” The text prompt is kept fixed in all experiments for consistency in image-text alignment.

To ensure valid few-shot evaluation, we build class-balanced training sets with 1, 2, 4, 8, 16, and 32 shots per class. We enforce strict WSI-level separation and patient-level isolation. All training patches come only from training WSIs. The validation set uses only WSIs that never appear in training. For both datasets, we used the same WSI-level split strategy: 2 WSIs for training, 2 WSIs for validation, and all remaining WSIs for testing. The validation set therefore contains 136 samples for Vanderbilt and 790 samples for Cornell. The few-shot training set contains 1, 2, 4, 8, 16, or 32 samples per class, corresponding to 5, 10, 20, 40, 80, or 160 total training samples per run, respectively. For each of the 10 independent runs, we constructed a distinct class-balanced training set from the same training WSIs, while keeping the validation and test WSIs fixed. For each run and class, we first form a 32-sample superset from the training WSIs. We fix a per-run permutation and derive smaller shots as prefixes of this order, where each smaller-shot subset is fully contained within all larger-shot sets. This design guarantees that, for example, the 4-shot configuration contains the exact images from the

Model	Method	Accuracy						
		0	1	2	4	8	16	32
CLIP Radford et al. (2021)	Adapter Gao et al. (2024)	0.3352±0.0000	0.3873±0.1108	0.2960±0.0881	0.4290±0.1478	0.5124±0.0656	0.5617±0.0385	0.5947±0.0689
	Classifier Ioffe and Szegedy (2015)	0.3352±0.0000	0.1200±0.0910	0.2838±0.1684	0.4445±0.1244	0.4997±0.1906	0.6686±0.0722	0.7140±0.0317
	LoRA Hu et al. (2022)	0.3352±0.0000	0.3671±0.1132	0.4430±0.1116	0.4930±0.0829	0.5845±0.0668	0.6505±0.0462	0.7028±0.0655
	Vanilla Srinivasan et al. (2024)	0.3352±0.0000	0.3505±0.1088	0.4584±0.1421	0.5344±0.1184	0.6441±0.0728	0.6634±0.0709	0.7192±0.0392
CONCH Lu et al. (2024)	Adapter Gao et al. (2024)	0.0363±0.0000	0.0403±0.0022	0.0929±0.0536	0.4446±0.2094	0.6617±0.0381	0.7437±0.0215	0.8082±0.0224
	Classifier Ioffe and Szegedy (2015)	0.0363±0.0000	0.3468±0.1912	0.4802±0.1855	0.5584±0.2574	0.4664±0.2487	0.7222±0.0670	0.7892±0.0368
	LoRA Hu et al. (2022)	0.0363±0.0000	0.0379±0.0009	0.0390±0.0010	0.0486±0.0028	0.2806±0.0567	0.6002±0.0452	0.7251±0.0198
	Vanilla Srinivasan et al. (2024)	0.0363±0.0000	0.3045±0.1044	0.4764±0.1280	0.6060 ±0.1045	0.6929 ±0.0409	0.7707 ±0.0400	0.8098 ±0.0160
PLIP Huang et al. (2023)	Adapter Gao et al. (2024)	0.3283±0.0000	0.3615±0.0669	0.4200±0.1471	0.5659±0.0702	0.5707±0.0544	0.6034±0.0577	0.6361±0.0670
	Classifier Ioffe and Szegedy (2015)	0.3283±0.0000	0.4357 ±0.1658	0.4482±0.1909	0.5254±0.1531	0.6170±0.1088	0.6599±0.0693	0.6603±0.0683
	LoRA Hu et al. (2022)	0.3283±0.0000	0.1894±0.0739	0.2788±0.0874	0.2807±0.0737	0.4696±0.0876	0.5528±0.0841	0.5899±0.0658
	Vanilla Srinivasan et al. (2024)	0.3283±0.0000	0.3038±0.1518	0.4901 ±0.1743	0.5830±0.1387	0.5719±0.1387	0.6184±0.1013	0.6256±0.0622
ResNet-18 He et al. (2016)	Supervised Learning	0.1047±0.0000	0.1078±0.0016	0.1345±0.0014	0.3288±0.0834	0.3466±0.1019	0.2693±0.0868	0.1159±0.0448
SSL Ciga et al. (2022)	Supervised Learning	0.3491 ±0.0000	0.3479±0.0003	0.1523±0.1667	0.2888±0.1584	0.2899±0.1851	0.2968±0.0874	0.1668±0.1116

Model	Method	AUC						
		0	1	2	4	8	16	32
CLIP Radford et al. (2021)	Adapter Gao et al. (2024)	0.4462±0.0000	0.5518±0.0599	0.5766±0.0478	0.6145±0.0271	0.6645±0.0133	0.7075±0.0324	0.7317±0.0414
	Classifier Ioffe and Szegedy (2015)	0.4462±0.0000	0.6344±0.0487	0.6699±0.0309	0.7171±0.0273	0.7452±0.0291	0.7867±0.0159	0.8064±0.0153
	LoRA Hu et al. (2022)	0.4462±0.0000	0.5476±0.0715	0.5969±0.0434	0.6284±0.0274	0.7190±0.0176	0.7849±0.0159	0.8148±0.0117
	Vanilla Srinivasan et al. (2024)	0.4462±0.0000	0.6132±0.0534	0.6399±0.0553	0.6824±0.0308	0.7463±0.0161	0.7909±0.0189	0.8245±0.0135
CONCH Lu et al. (2024)	Adapter Gao et al. (2024)	0.4400±0.0000	0.4908±0.0066	0.5577±0.0515	0.6727±0.0911	0.7626±0.0143	0.8078±0.0100	0.8556±0.0110
	Classifier Ioffe and Szegedy (2015)	0.4400±0.0000	0.7525 ±0.0390	0.7833 ±0.0325	0.8016 ±0.0434	0.8195 ±0.0281	0.8502 ±0.0114	0.8700 ±0.0093
	LoRA Hu et al. (2022)	0.4400±0.0000	0.4877±0.0039	0.4939±0.0097	0.5028±0.0128	0.6395±0.0142	0.7381±0.0135	0.7954±0.0099
	Vanilla Srinivasan et al. (2024)	0.4400±0.0000	0.6424±0.0476	0.7113±0.0331	0.7737±0.0239	0.7789±0.0182	0.8294±0.0269	0.8573±0.0113
PLIP Huang et al. (2023)	Adapter Gao et al. (2024)	0.4825 ±0.0000	0.6349±0.0208	0.6120±0.0731	0.6588±0.0196	0.7060±0.0186	0.7377±0.0234	0.7685±0.0068
	Classifier Ioffe and Szegedy (2015)	0.4825 ±0.0000	0.6719±0.0496	0.6775±0.0368	0.7239±0.0332	0.7419±0.0230	0.7752±0.0203	0.7796±0.0146
	LoRA Hu et al. (2022)	0.4825 ±0.0000	0.6171±0.0251	0.6298±0.0454	0.6674±0.0138	0.7109±0.0168	0.7755±0.0163	0.8198±0.0087
	Vanilla Srinivasan et al. (2024)	0.4825 ±0.0000	0.6425±0.0359	0.6721±0.0512	0.7055±0.0256	0.7311±0.0176	0.7653±0.0170	0.8102±0.0089
ResNet-18 He et al. (2016)	Supervised Learning	0.4541±0.0000	0.4544±0.0003	0.5437±0.0004	0.5007±0.0356	0.5092±0.0349	0.5014±0.0569	0.4985±0.0484
SSL Ciga et al. (2022)	Supervised Learning	0.4657±0.0000	0.4645±0.0008	0.5832±0.0895	0.4632±0.0475	0.5170±0.0850	0.5112±0.0729	0.5465±0.0472

Model	Method	F1						
		0	1	2	4	8	16	32
CLIP Radford et al. (2021)	Adapter Gao et al. (2024)	0.1080±0.0000	0.1902±0.0595	0.1743±0.0497	0.2425±0.0692	0.2992±0.0238	0.3347±0.0153	0.3495±0.0356
	Classifier Ioffe and Szegedy (2015)	0.1080±0.0000	0.0722±0.0489	0.1797±0.0930	0.2729±0.0613	0.3080±0.1029	0.3991±0.0339	0.4263±0.0145
	LoRA Hu et al. (2022)	0.1080±0.0000	0.1738±0.0511	0.2111±0.0526	0.2627±0.0494	0.3505±0.0187	0.4013±0.0118	0.4208±0.0313
	Vanilla Srinivasan et al. (2024)	0.1080±0.0000	0.2108±0.0464	0.2480±0.0695	0.3068±0.0424	0.3738±0.0225	0.4055±0.0238	0.4328±0.0200
CONCH Lu et al. (2024)	Adapter Gao et al. (2024)	0.0376±0.0000	0.0439±0.0036	0.0856±0.0404	0.2755±0.1189	0.3892±0.0177	0.4403±0.0120	0.4821±0.0192
	Classifier Ioffe and Szegedy (2015)	0.0376±0.0000	0.2117±0.1053	0.3124 ±0.1029	0.3248±0.1704	0.3088±0.1373	0.4590±0.0319	0.4915±0.0291
	LoRA Hu et al. (2022)	0.0376±0.0000	0.0409±0.0013	0.0421±0.0015	0.0509±0.0025	0.2050±0.0250	0.3619±0.0257	0.4331±0.0113
	Vanilla Srinivasan et al. (2024)	0.0376±0.0000	0.2086±0.0464	0.2981±0.0595	0.3799 ±0.0387	0.4109 ±0.0275	0.4640 ±0.0283	0.4957 ±0.0197
PLIP Huang et al. (2023)	Adapter Gao et al. (2024)	0.1492 ±0.0000	0.2317±0.0352	0.2482±0.0829	0.3224±0.0318	0.3408±0.0219	0.3596±0.0222	0.3852±0.0260
	Classifier Ioffe and Szegedy (2015)	0.1492 ±0.0000	0.2628 ±0.0658	0.2726±0.0766	0.3251±0.0644	0.3698±0.0465	0.3986±0.0280	0.4057±0.0218
	LoRA Hu et al. (2022)	0.1492 ±0.0000	0.1409±0.0364	0.1964±0.0368	0.2154±0.0305	0.3161±0.0311	0.3682±0.0415	0.4038±0.0273
	Vanilla Srinivasan et al. (2024)	0.1492 ±0.0000	0.2053±0.0781	0.2815±0.0836	0.3423±0.0539	0.3489±0.0472	0.3809±0.0402	0.4161±0.0233
ResNet-18 He et al. (2016)	Supervised Learning	0.0679±0.0000	0.0693±0.0007	0.0746±0.0013	0.1165±0.0366	0.1576±0.0370	0.1289±0.0321	0.0996±0.0321
SSL Ciga et al. (2022)	Supervised Learning	0.1210±0.0000	0.1221±0.0001	0.0556±0.0584	0.1233±0.0455	0.1564±0.0844	0.1204±0.0229	0.1002±0.0505

Table 1: **Few-shot classification performance on Cornell.** Mean±SD Accuracy, AUC, and F1 across shot settings {0, 1, 2, 4, 8, 16, 32} for all backbone and adaptation pairs. Here, k -shot denotes k labeled images per class used for fine-tuning. For the VLMs, 0-shot denotes inference without few-shot adaptation; for the ResNet-18 linear-probe baselines, 0-shot denotes the reference result obtained with a randomly initialized linear classification head before any few-shot training. Bold indicates the best value within each shot column (higher is better).

2-shot set plus two additional images per class, enabling a controlled comparison across shot levels. We also maximize diversity across runs, especially for small shots. Each run receives a distinct round-robin slice of the per-class pool, and we prioritize that slice to reduce overlap across runs. All sampling uses fixed random seeds and deterministic rules.

4.3 Experimental Design

To ensure robust and stable model optimization, we performed a systematic exploration of key hyperparameters. Training was conducted using the AdamW optimizer with an initial learning rate of 3×10^{-4} , incorporating a linear warm-up stage followed by scheduled decay to ensure smooth and stable convergence. To further strengthen generalization and mitigate overfitting, we uniformly applied data augmentation using the Albumentations Buslaev et al. (2020) library,

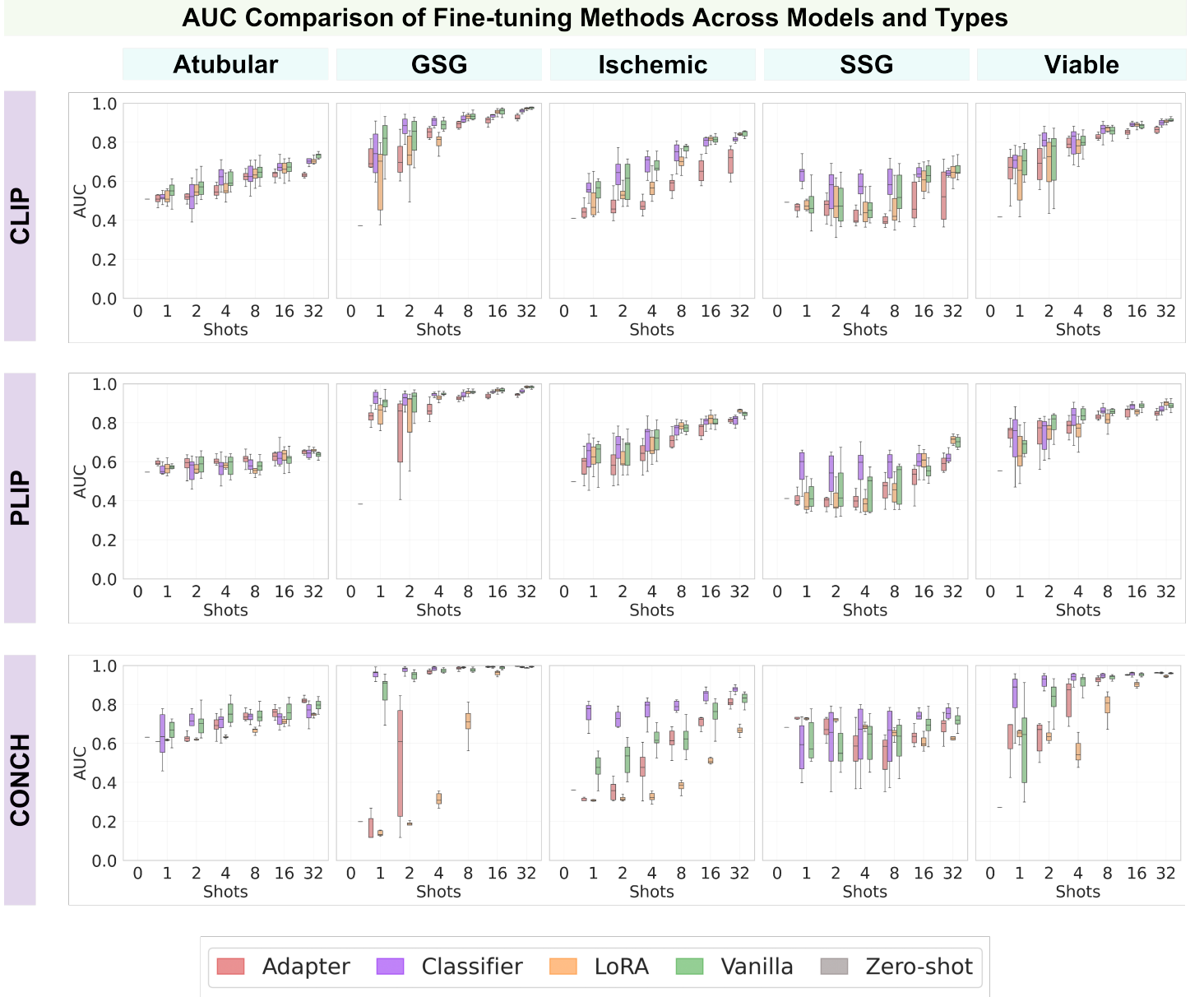


Figure 5: **Discrimination across subtypes on Cornell.** This figure summarizes per-class discrimination performance across 10 runs. For each glomerular subtype, we report the distribution of AUC across VLM backbones, fine-tuning strategies, and shot settings. GSG and SSG denote Globally Sclerotic Glomerulus and Segmentally Sclerotic Glomerulus, respectively.

including vertical and horizontal flips, transpose, random 90° rotation, blur-based perturbations (MotionBlur, MedianBlur, Blur), and contrast/texture adjustments (CLAHE, RandomBrightnessContrast, Sharpen), and employed early stopping with consistent patience and minimum-delta settings across all experiments. For parameter-efficient adaptation, we set LoRA rank to $r = 16$ and scaling factor to $\alpha = 32$, and set the adapter bottleneck hidden dimension to $m = 64$. For Classifier tuning, we used an MLP BatchNorm head with one hidden layer; the input feature dimension is 512. We optimized only the classifier-head parameters using AdamW with a learning rate of 1×10^{-3} and weight decay of 1×10^{-4} ; the hidden dimension was set to 512,

and the dropout rate was set to 0.1. The ResNet-18 and SSL baselines use the same WSI-level splits and few-shot budgets as the VLM experiments. All experiments were conducted on an NVIDIA RTX 6000 Ada Generation with 48 GB of VRAM. For quantitative results, we report the mean and standard deviation over 10 independent runs. Uniform Manifold Approximation and Projection (UMAP) visualizations are generated from the run achieving the highest AUC.

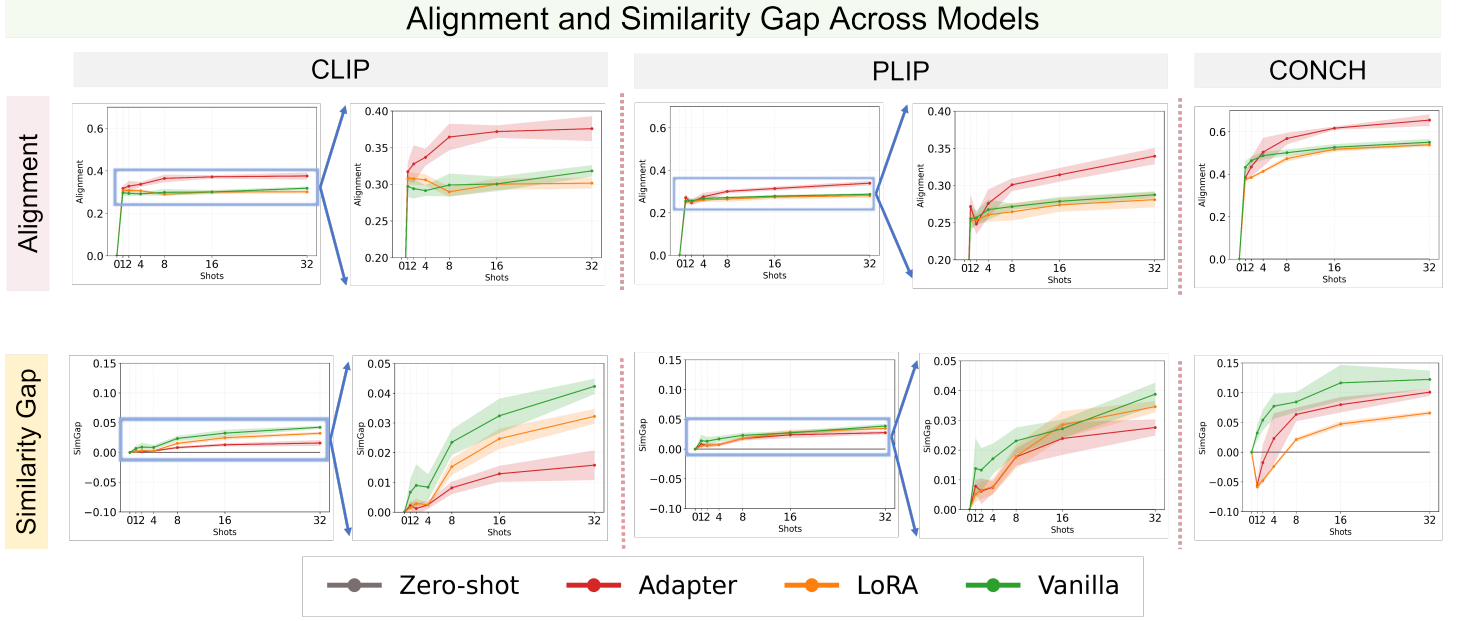


Figure 6: **Alignment and similarity gap across models on Cornell.** This figure shows shot-wise trajectories of Alignment and Similarity Gap for CLIP, PLIP, and CONCH under different fine-tuning strategies. Similarity Gap is defined as mean positive similarity (Alignment) minus mean negative similarity. CLIP and PLIP are demonstrated in both regular and enlarged panels to show their behavior.

4.4 Evaluation Metrics

We evaluated all methods on a five-class glomerular classification task using a fixed train-validation split and standardized class prompts. Performance was reported using overall accuracy, macro-AUC, and F1 score, with identical data augmentation and early stopping applied to ensure a fair comparison Wang and Carvalho (2024).

In addition to standard metrics, we further assessed internal representation alignment by measuring the alignment and similarity gap, as depicted in Fig. 3. Alignment quantifies the average cosine similarity between paired image-text embeddings, reflecting how well the two modalities are semantically matched, and is computed in Eq. 2:

$$\text{Alignment} = \frac{1}{N} \sum_{i=1}^N \cos(\mathbf{z}_i, \mathbf{w}_{y_i}). \quad (2)$$

The similarity gap measures the mean separation between the positive class prompt and the negative class prompts, capturing the model’s discriminative ability in distinguishing aligned from unaligned class-level representations and is defined in Eq. 3:

$$\begin{aligned} \text{Similarity Gap} = & \frac{1}{N} \sum_{i=1}^N \cos(\mathbf{z}_i, \mathbf{w}_{y_i}) \\ & - \frac{1}{N(C-1)} \sum_{i=1}^N \sum_{\substack{c=1 \\ c \neq y_i}}^C \cos(\mathbf{z}_i, \mathbf{w}_c). \end{aligned} \quad (3)$$

where $\mathbf{z}_i$ denotes the normalized image embedding of the i -th sample, y_i is its class label, C is the number of classes, and $\mathbf{w}_c$ denotes the normalized text embedding of the prompt for class c . The negative term averages over all class prompts other than the ground-truth class prompt for each image. Therefore, the similarity gap increases when positive class-prompt similarity rises, negative class-prompt similarity decreases, or both. Together, these two metrics provide complementary perspectives on multimodal consistency. We also visualized multimodal feature distributions using UMAP for dimensionality reduction and kernel density estimation (KDE) for density-based visualization to provide qualitative insights. UMAP was chosen as it provides a compact yet faithful two-dimensional representation of the embedding space, facilitating visual comparison of inter-class and cross-modal distributions. All UMAP embeddings and KDE visualizations were produced using consistent settings across experiments.

Beyond alignment-based measures, we report Intra-Class Distance and Silhouette Score. For intra-class distance, let $\mathbf{x}_i$ denote the image representation used for geometric analysis and y_i its class label. For Classifier tuning, $\mathbf{x}_i$ is the penultimate hidden representation of the tuned MLP head fed by fixed 512-D CONCH features; for the other fine-tuning strategies, $\mathbf{x}_i$ is the corresponding image embedding produced by the tuned encoder. We obtain the two-dimensional UMAP projection $\mathbf{u}_i = \text{UMAP}(\mathbf{x}_i)$ using fixed hyperparameters and seed across experiments. Intra-class distance is the mean pairwise euclidean distance

among same-class samples in the UMAP space, as shown in Eq. 4:

$$\text{ICD} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{2}{n_c(n_c - 1)} \sum_{\substack{i < j \\ y_i = y_j = c}} \|\mathbf{u}_i - \mathbf{u}_j\|_2. \quad (4)$$

where $\mathcal{C}$ is the set of classes and n_c is the number of samples in class c . To assess separability in the same representation space, we compute the Silhouette score Rousseeuw (1987) on $\{\mathbf{x}_i\}$ with cosine distance. For each sample, a_i denotes its mean distance to samples in the same class, whereas b_i denotes its minimum mean distance to any other class. The final score is

$$S = \frac{1}{N} \sum_{i=1}^N \frac{b_i - a_i}{\max\{a_i, b_i\}}, \quad S \in [-1, 1]. \quad (5)$$

Detailed definitions of a_i and b_i are provided in Appendix A.

To quantify model confidence and better characterize decision boundaries, we additionally evaluate ECE Guo et al. (2017) and boundary margin, defined as the top-1 vs. top-2 probability gap. We compute top-1 ECE using $M = 15$ equal-width confidence bins over $[0, 1]$:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|. \quad (6)$$

Here, N is the number of test samples and $|B_m|$ is the number of samples in bin B_m ; lower ECE indicates better-calibrated predicted probabilities. The bin-level accuracy and confidence definitions are provided in Appendix A.

We quantify boundary margin as a simple proxy for decision boundary sharpness in the model's predicted probability space. For each sample i with predicted class probabilities $\mathbf{p}_i = \text{softmax}(\mathbf{s}_i)$, let $p_i^{(1)}$ and $p_i^{(2)}$ denote the largest and second-largest entries in $\mathbf{p}_i$, respectively. The boundary margin is defined in Eq. 7:

$$\Delta p_i = p_i^{(1)} - p_i^{(2)}. \quad (7)$$

Larger Δp_i indicates a clearer separation between the top prediction and the closest alternative, suggesting that the sample lies farther from the decision boundary in the model's predicted probability space.

For clinical label-quality assessment, we additionally performed an expert label audit on a balanced subset of 500 Cornell patches, with 100 patches sampled from each of the five classes. Three doctors independently reviewed the class assignment of each patch using the same five-class schema. We quantified inter-rater reliability using Fleiss' κ Fleiss (1971), which measures agreement beyond chance. Its final form is

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \quad (8)$$

where $\bar{P}$ is the mean observed agreement and $\bar{P}_e$ is the expected agreement by chance. Their detailed computation is provided in Appendix A. In our study, $N = 500$, $n = 3$, and $k = 5$. We computed physician recall within each ground-truth class and a one-vs-rest Fleiss' κ for each class by collapsing that class against the remaining four classes. The Average entry in the class-wise κ row is therefore a macro-average across the five one-vs-rest Fleiss' κ values.

5. Results

5.1 Model Validation

5.1.1 Classification Performance

Tables 1 and S1 show a shared supervision trend across Cornell and Vanderbilt datasets: performance improves as shots increase, and AUC changes are the most consistent among the three metrics. In both ROC-AUC plots from Figs. 4 and S1, curves progressively shift toward the upper-left region. At higher shots, the strongest configurations are dominated by CONCH-based models, with CONCH with Vanilla leading most high-shot endpoints and CONCH with Classifier giving the top AUC on Cornell.

Across backbones, CONCH shows the strongest medium-to-high-shot behavior on both datasets. Cornell presents a low-shot regime where PLIP is competitive, but CONCH becomes dominant from moderate supervision onward. On the Vanderbilt dataset, CONCH remains the most competitive backbone through most shot settings, while CLIP and PLIP improve more gradually and stay behind at high shots. Across adaptation strategies, Vanilla exhibits the most persistent cumulative gains and most frequently attains the best medium/high-shot results, especially for Accuracy and F1. Adapter and LoRA are generally intermediate, with LoRA often narrowing the gap at higher shots. Classifier tuning is the most dataset-dependent strategy: it is strong for CONCH-AUC on Cornell, but it is consistently weak on the Vanderbilt dataset across backbones. The clearest cross-dataset split is therefore observed in Classifier tuning, which performs competitively on Cornell but degrades on the Vanderbilt dataset. A second split is in improvement breadth: Cornell shows broader gains across multiple backbone-strategy combinations, whereas the Vanderbilt dataset concentrates its strongest gains in CONCH with Vanilla adaptation. Qualitatively, the Cornell and Vanderbilt boxplots show a shared supervision trend: as shots increase, AUC medians generally move upward and the distributions become more concentrated, with Vanilla frequently yielding the strongest high-shot profiles (Figs. 5 and S2). In Cornell, method trajectories are comparatively aligned, and backbone-level differences are present but moderate. By class, GSG and Viable tend to show higher and tighter distributions, whereas Ischemic and SSG remain broader with

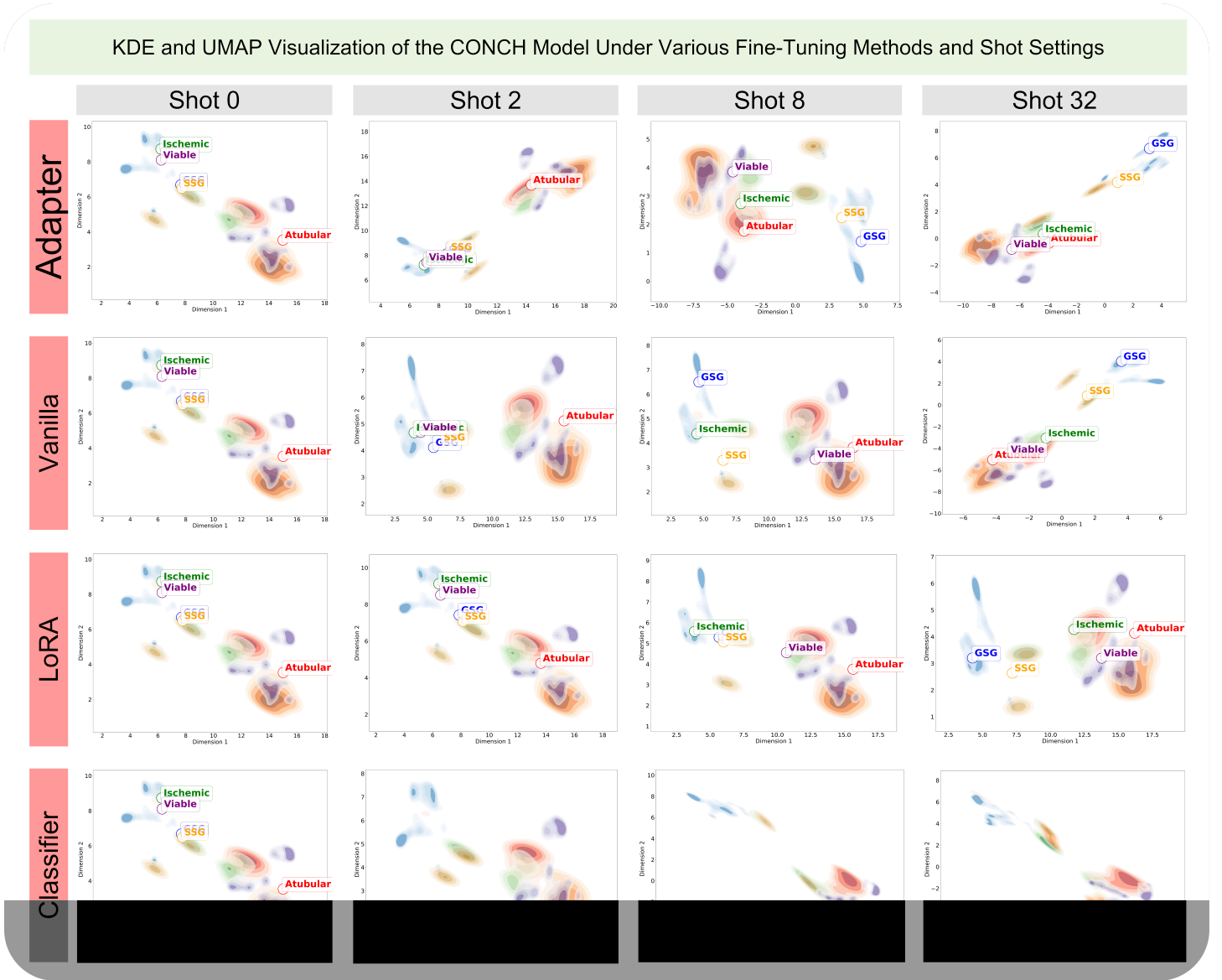


Figure 7: **UMAP/KDE visualization of embeddings across shots on Cornell.** This figure shows text and image embeddings from the best CONCH run across shot settings. For Classifier tuning, text is not updated and only image-side embeddings are shown. Specifically, for Classifier tuning we visualize the penultimate hidden representation of the tuned MLP head (fed by fixed 512-D CONCH features), which changes as the head is updated. The companion Table 2 reports Intra-Class Distance and Silhouette Score.

longer whiskers, indicating greater classification difficulty.

In the Vanderbilt dataset, the most salient pattern is stronger separation across adaptation strategies. Vanilla, and in several settings LoRA, show clearer upward median shifts and tighter high-shot boxes, while Classifier tuning remains consistently lower and more dispersed across backbones and classes. Relative to Cornell, this method-level divergence is substantially larger in the Vanderbilt dataset and becomes the dominant qualitative feature of the boxplot results.

5.1.2 Cross-Multimodalities Comparison

As shown in Figs. 6 and S3, Cornell and Vanderbilt datasets exhibit the same directional pattern in cross-modal metrics. With increasing supervision, Alignment generally rises, and Similarity Gap shifts from near-zero to clearly positive values. Across backbones, Adapter most often gives the highest Alignment, whereas Vanilla most often gives the largest Similarity Gap, with LoRA typically in between.

The main cross-dataset difference is therefore the strength of these trends. Cornell shows steeper increases and clearer method separation, especially at medium-to-high shots and most visibly on CONCH. Vanderbilt follows the same ordering but with flatter trajectories and more overlap across

Model	Method	Intra-Class Distance							Silhouette Score						
		0	1	2	4	8	16	32	0	1	2	4	8	16	32
CLIP Radford et al. (2021)	Adapter Gao et al. (2024)	2.5430	2.4421	2.4591	2.5170	3.1200	2.6249	2.4688	0.0373	0.0330	0.0318	0.0307	0.0701	0.0677	0.0803
	Classifier Ioffe and Szegedy (2015)	2.5430	2.4863	2.3897	2.6498	2.7732	2.8545	2.9882	0.0373	0.0490	0.0292	0.0718	0.0884	0.0807	0.0968
	LoRA Hu et al. (2022)	2.5430	2.3812	2.4981	2.5857	3.0166	2.5515	2.4130	0.0373	0.0332	0.0310	0.0331	0.0591	0.0557	0.0968
	Vanilla Srinivasan et al. (2024)	2.5430	2.6376	2.3177	2.3004	2.5726	2.3463	2.6607	0.0373	0.0440	0.0279	0.0410	0.0773	0.0665	0.1252
CONCH Lu et al. (2024)	Adapter Gao et al. (2024)	3.0591	3.1978	3.4481	3.1913	3.0749	3.0657	2.8749	0.1812	0.1794	0.1729	0.1614	0.1931	0.2150	0.3158
	Classifier Ioffe and Szegedy (2015)	3.0591	2.9927	3.0976	2.8299	3.6277	3.6330	3.6131	0.1812	0.1870	0.2002	0.1775	0.2929	0.2301	0.2887
	LoRA Hu et al. (2022)	3.0591	3.1626	3.0679	3.1751	3.2274	3.1471	3.0469	0.1812	0.1805	0.1806	0.1800	0.1774	0.1767	0.1873
	Vanilla Srinivasan et al. (2024)	3.0591	3.3166	3.2381	3.0948	3.1859	2.8229	2.3254	0.1812	0.1864	0.1792	0.2041	0.3027	0.3453	0.3923
PLIP Huang et al. (2023)	Adapter Gao et al. (2024)	4.6734	4.7702	5.3012	4.5804	4.6559	4.5383	4.3455	0.0768	0.0801	0.0768	0.0677	0.1049	0.0900	0.0975
	Classifier Ioffe and Szegedy (2015)	4.6734	4.0596	5.0829	4.5074	4.2935	4.4788	4.4356	0.0768	0.1336	0.1356	0.1350	0.1314	0.1450	0.1548
	LoRA Hu et al. (2022)	4.6734	4.8622	4.3826	5.0877	4.7062	4.7701	5.0552	0.0768	0.0711	0.0703	0.0687	0.0915	0.0964	0.1260
	Vanilla Srinivasan et al. (2024)	4.6734	4.6071	4.6863	5.1612	4.7397	4.8918	4.9956	0.0768	0.0790	0.0762	0.0862	0.0951	0.1109	0.1214

Table 2: **Intra-Class Distance and Silhouette Score on Cornell.** Few-shot geometric metrics under three backbones: CLIP, CONCH, and PLIP across shot settings $\{0, 1, 2, 4, 8, 16, 32\}$ and four fine-tuning strategies. Intra-Class Distance measures within-class compactness in the UMAP space (lower is better), and Silhouette Score measures cluster separability in the original embedding space using cosine distance (higher is better). For Classifier tuning, both metrics are computed from the penultimate representation of the tuned MLP head.

methods in several settings. In short, the two datasets share a consistent trend structure, while differing in effect magnitude.

Consistent with the alignment/similarity trends, the UMAP projections (Figs. 7 and S4) show progressively improved structure with increasing shots: image clusters become more compact, class-wise regions become more separable, and text anchors move closer to their corresponding image groups in the stronger settings. On Cornell, this refinement is most evident for CONCH with Vanilla, where high-shot embeddings form the clearest modality-consistent organization. On Vanderbilt, the same directional improvement is present, but Adapter and Vanilla remain closer to each other in visual quality, with less pronounced separation between the two.

The quantitative geometry metrics follow the same pattern (Tables 2 and S3). In Cornell, high-shot CONCH with Vanilla most clearly combines lower Intra-Class Distance with higher Silhouette, matching the cleaner UMAP layout. In Vanderbilt, both Adapter and Vanilla improve the geometry, but their gap is modest, again reflecting weaker method separation than in Cornell. Across both datasets, Classifier tends to show less coherent geometric behavior, which is also visible as less stable cluster structure in UMAP.

5.1.3 Confidence Calibration and Decision Boundary Margin Analysis

We additionally report ECE and boundary margin (Tables 3 and S2) to characterize confidence quality and decision boundary behavior. Lower ECE indicates better calibration, while higher margin indicates clearer class separation.

On Cornell, metrics are broadly aligned: increasing supervision generally improves classification quality, calibration, and margin in stronger settings, particularly for high-shot CONCH and Vanilla adaptation, whereas the extreme low-shot regime shows the main discrepancy, with

some settings remaining low-performing despite producing high-confidence and poorly calibrated predictions, as indicated by ECE.

On Vanderbilt, the trend is more heterogeneous across methods. Although classification improves at higher shots for the better adaptation strategies, ECE often remains comparatively high, and method-level differences become more pronounced with supervision. The clearest mismatch is concentrated in Classifier tuning, where confidence-related signals remain strong while classification quality stays weak across multiple backbones.

Cross-dataset divergence is therefore most evident along two axes: strategy and shot regime. Strategy-wise, Classifier tuning is the largest split between datasets; shot-wise, Cornell mismatches are mainly early-shot, whereas Vanderbilt mismatches are more persistent in medium-to-high-shot settings.

5.2 Clinical Validation

5.2.1 Expert Label Audit

To assess label reliability from a clinical perspective, we conducted an expert label audit on a balanced subset of 500 Cornell patches (100 per class). The class-wise summary is reported in Table 4. When agreement was computed jointly across all 500 reviewed patches and all five classes, the overall multiclass Fleiss' κ was 0.762, indicating substantial agreement. This overall value is distinct from the table's Average one-vs-rest κ entry (0.7615), which is a macro-average across classes rather than a single multiclass agreement estimate. Among the 500 reviewed patches, 365 (73.0%) received identical labels from all three raters, and 489 (97.8%) received the same label from at least two raters.

Agreement was not uniform across classes. Global Glomerulosclerosis showed near-perfect inter-rater agree-

Model	Method	ECE (Top-1)						
		0	1	2	4	8	16	32
CLIP Radford et al. (2021)	Shots							
	Adapter Gao et al. (2024)	0.1220±0.0000	0.1336±0.0583	0.0832±0.0511	0.1926±0.0830	0.1784±0.0450	0.1460±0.0370	0.1307±0.0490
	Classifier Ioffe and Szegedy (2015)	0.1220±0.0000	0.1317±0.0625	0.1067±0.0370	0.1411±0.0984	0.0932±0.0471	0.0618±0.0312	0.0476±0.0219
	LoRA Hu et al. (2022)	0.1220±0.0000	0.1119±0.0523	0.1631±0.0520	0.2231±0.0716	0.0798±0.0271	0.0641±0.0271	0.0525±0.0428
CONCH Lu et al. (2024)	Vanilla Srinivasan et al. (2024)	0.1220±0.0000	0.1702±0.0726	0.1150±0.0848	0.1597±0.0813	0.0741±0.0377	0.0822±0.0489	0.0829±0.0341
	Adapter Gao et al. (2024)	0.1790±0.0000	0.8659±0.0064	0.7665±0.0889	0.3776±0.2461	0.1841±0.0353	0.1462±0.0238	0.1234±0.0178
	Classifier Ioffe and Szegedy (2015)	0.1790±0.0000	0.1875±0.1001	0.2588±0.1676	0.3281±0.1150	0.2727±0.1230	0.3610±0.1562	0.2004±0.1815
	LoRA Hu et al. (2022)	0.1790±0.0000	0.8672±0.0013	0.8657±0.0021	0.8349±0.0121	0.5096±0.0671	0.2203±0.0330	0.1575±0.0167
PLIP Huang et al. (2023)	Vanilla Srinivasan et al. (2024)	0.1790±0.0000	0.4287±0.1241	0.2662±0.1084	0.1938±0.0695	0.1255±0.0405	0.1104±0.0262	0.0954±0.0192
	Adapter Gao et al. (2024)	0.1120±0.0000	0.1102±0.0501	0.1022±0.0888	0.1736±0.0550	0.0658±0.0203	0.0763±0.0423	0.0582±0.0383
	Classifier Ioffe and Szegedy (2015)	0.1120±0.0000	0.2462±0.0928	0.2587±0.1440	0.1898±0.1135	0.1523±0.1122	0.0826±0.0540	0.0780±0.0514
	LoRA Hu et al. (2022)	0.1120±0.0000	0.3033±0.0983	0.1900±0.0762	0.1710±0.0563	0.1203±0.0584	0.1388±0.0824	0.1407±0.0672
ResNet-18 He et al. (2016)	Vanilla Srinivasan et al. (2024)	0.1120±0.0000	0.3047±0.1390	0.1433±0.0988	0.1161±0.0988	0.1347±0.0901	0.1023±0.0794	0.1226±0.0622
	Supervised Learning	0.2614±0.0000	0.2523±0.0015	0.2496±0.0025	0.1039±0.0364	0.1005±0.0359	0.0889±0.0678	0.1655±0.0494
SSL Ciga et al. (2022)	Supervised Learning	0.0877±0.0000	0.0896±0.0004	0.2444±0.0280	0.1550±0.0519	0.1619±0.1059	0.1056±0.0441	0.1104±0.0765

Model	Method	Boundary Margin (Mean)						
		0	1	2	4	8	16	32
CLIP Radford et al. (2021)	Shots							
	Adapter Gao et al. (2024)	0.0006±0.0000	0.0589±0.0185	0.0459±0.0220	0.0525±0.0197	0.1491±0.0294	0.2230±0.0484	0.2762±0.1015
	Classifier Ioffe and Szegedy (2015)	0.0006±0.0000	0.0181±0.0156	0.1023±0.1293	0.1488±0.1683	0.3085±0.2280	0.5333±0.1413	0.5879±0.0519
	LoRA Hu et al. (2022)	0.0006±0.0000	0.0642±0.0247	0.0773±0.0342	0.0520±0.0132	0.3226±0.0685	0.4614±0.0681	0.5753±0.0503
CONCH Lu et al. (2024)	Vanilla Srinivasan et al. (2024)	0.0006±0.0000	0.2218±0.1082	0.2319±0.1293	0.1751±0.0861	0.4876±0.0757	0.5545±0.0902	0.6614±0.0283
	Adapter Gao et al. (2024)	0.0114±0.0000	0.8240±0.0147	0.7493±0.0601	0.6818±0.0787	0.7192±0.0404	0.7987±0.0322	0.8731±0.0145
	Classifier Ioffe and Szegedy (2015)	0.0114±0.0000	0.0191±0.0208	0.0191±0.0164	0.0681±0.0640	0.0503±0.1048	0.1979±0.2749	0.5042±0.2995
	LoRA Hu et al. (2022)	0.0114±0.0000	0.8208±0.0018	0.8201±0.0036	0.7858±0.0168	0.6286±0.0259	0.6797±0.0277	0.7864±0.0144
PLIP Huang et al. (2023)	Vanilla Srinivasan et al. (2024)	0.0114±0.0000	0.5468±0.0561	0.5566±0.0605	0.6510±0.0941	0.6797±0.0552	0.7877±0.0592	0.8298±0.0438
	Adapter Gao et al. (2024)	0.0016±0.0000	0.1646±0.0216	0.1466±0.0315	0.1418±0.0268	0.3196±0.0546	0.4138±0.0739	0.4688±0.0348
	Classifier Ioffe and Szegedy (2015)	0.0016±0.0000	0.1031±0.0937	0.1439±0.1589	0.2292±0.2142	0.3272±0.1777	0.4813±0.1514	0.4694±0.1190
	LoRA Hu et al. (2022)	0.0016±0.0000	0.2311±0.0583	0.1814±0.0509	0.1813±0.0398	0.3407±0.0591	0.4850±0.0384	0.5315±0.0222
ResNet-18 He et al. (2016)	Vanilla Srinivasan et al. (2024)	0.0016±0.0000	0.3785±0.1417	0.3038±0.0939	0.3535±0.0609	0.4180±0.0727	0.4564±0.0483	0.5517±0.0525
	Supervised Learning	0.0914±0.0000	0.0885±0.0004	0.1355±0.0016	0.1355±0.0350	0.0733±0.0142	0.1057±0.0397	0.0521±0.0057
SSL Ciga et al. (2022)	Supervised Learning	0.0481±0.0000	0.0475±0.0002	0.0896±0.0341	0.0266±0.0056	0.0213±0.0047	0.0358±0.0165	0.0244±0.0070

Table 3: **ECE and boundary margin on Cornell.** ECE (Top-1) and boundary margin mean scores (Mean±SD) across shots {0, 1, 2, 4, 8, 16, 32}. Column-wise best values are bolded (lower is better for ECE; higher is better for boundary margin).

ment, while Ischemic Glomeruli showed the lowest agreement. Viable Glomeruli had high class-wise physician recall/sensitivity but lower one-vs-rest inter-rater κ , indicating that per-class recognition and agreement do not always move together. Overall, the expert audit supports the reliability and clinical plausibility of the benchmark labels while also showing that some classes are more challenging than others. We focused this audit on the Cornell subset because Cornell serves as the primary fine-grained glomerular subtype benchmark in this study, and its five classes represent closely related morphologic subtypes with known boundary ambiguity, particularly across ischemic change, segmental sclerosis, and global sclerosis.

5.3 Discussion

Across both cohorts, increasing supervision consistently improves downstream discrimination. At the same time, the gain profile differs by dataset: Cornell shows steeper and more stable improvements, whereas Vanderbilt follows the same direction with smaller effect size and stronger method sensitivity. This overall pattern indicates that few-shot scaling is reproducible across centers, but its practical benefit still depends on backbone and adaptation choice

under dataset-specific distributions.

From the backbone perspective, pathology-specialized models outperform general CLIP, with CONCH showing the strongest high-shot ceiling. The advantage of CONCH/PLIP over CLIP is consistent with domain-informed pretraining: pathology-native representations start from a more relevant semantic prior and require fewer labels to reach useful discrimination. The further separation between CONCH and PLIP at higher shots likely reflects architectural and objective-level differences in multimodal pretraining, which become more visible once supervision is sufficient to exploit model capacity. In practical terms, this suggests a two-stage pattern in which domain knowledge dominates early-shot utility, while representational flexibility determines high-shot headroom.

From the adaptation perspective, Vanilla is the most consistently effective at medium-to-high shots, while Classifier tuning is the least robust. Vanilla fine-tuning can jointly reshape feature geometry and decision boundaries rather than only reweighting a fixed embedding space. Adapter and LoRA provide useful but more constrained updates, and they often improve progressively with additional labels. By contrast, Classifier tuning is highly conditional: it can appear competitive when frozen features are already well

Metric	Atubular Glomeruli	Global Glomerulosclerosis	Ischemic Glomeruli	Segmental Glomerulosclerosis	Viable Glomeruli	Average
Doctor 1	0.6900	1.0000	0.9700	0.8000	0.9500	0.8820
Doctor 2	0.6900	1.0000	0.4100	0.7600	0.9500	0.7620
Doctor 3	0.7300	1.0000	0.7100	0.9100	0.9000	0.8500
One-vs-rest Fleiss' κ	0.8419	0.9835	0.5497	0.7422	0.6900	0.7615

Table 4: **Human pathologist performance baseline on a blinded glomeruli subset.** Class-wise pathologist recall and inter-pathologist agreement are reported to contextualize model performance. Recall is computed within each ground-truth class. Fleiss' κ is computed per class using a one-vs-rest formulation; the Average entry for this row is the macro-average of the five one-vs-rest Fleiss' κ values.

aligned, but it degrades rapidly when representation shift is non-trivial, producing unstable behavior across datasets and shot regimes.

The cross-modal results add an important complementary view. Adapter tends to maximize alignment, whereas Vanilla more often enlarges the similarity gap and improves geometric separability. This difference is expected because alignment and similarity gap are complementary rather than interchangeable: higher alignment mainly reflects stronger positive-pair attraction, while a larger similarity gap requires simultaneous separation from mismatched classes. Accordingly, Adapter can lead alignment without always yielding the best classification, whereas Vanilla is more consistently associated with cleaner UMAP geometry (lower intra-class distance and higher silhouette in the strongest settings). Again, Cornell and Vanderbilt show the same direction but different intensity, with clearer method separation on Cornell.

Calibration and boundary analyses further refine this picture. The key failure mode is confident error, which is more persistent on Vanderbilt. High margin alone is not sufficient for reliability if calibration is poor. In this setting, confidence-related signals can remain strong while correctness stays low, creating a clinically relevant mismatch. In our results, this mismatch is concentrated in low-shot extremes on Cornell but persists into medium/high-shot settings on Vanderbilt for less robust strategies, indicating greater confidence fragility under external domain shift.

When model failure occurs, it is concentrated in two recurring scenarios. First, errors are most likely for glomeruli that lie along the morphological continuum from ischemic change to segmental sclerosis and then global sclerosis, where subtle structural differences make discrimination more difficult under a single-label five-class formulation. This interpretation is consistent with the broader and less stable class-wise behavior observed for the more difficult categories, particularly Ischemic and SSG. Second, failure also arises under external domain shift, especially on Vanderbilt and for weaker adaptation strategies such as Classifier tuning, where confidence-related signals can remain strong even when classification quality is limited. In this setting, the dominant risk is not simply low confidence, but confident

error, indicating that a high decision margin alone is insufficient evidence of reliable performance under cross-center variation.

These cross-center differences are consistent with, but cannot be attributed solely to, institution-specific appearance variation, including staining and scanner effects, superimposed on cohort heterogeneity. Within the two evaluated cohorts, increasing the target-cohort shot budget was associated with improved discrimination, embedding geometry, and cross-modal separation, particularly for stronger backbones and Vanilla adaptation. However, calibration fragility and method sensitivity persisted on Vanderbilt even at medium-to-high shot settings. Therefore, the present results suggest that additional in-domain supervision may help adaptation to center-specific visual variation, but they do not establish general mitigation of domain shift. Broader multi-center evaluations spanning additional scanners, staining protocols, and patient populations are needed to test this hypothesis.

The timing analysis adds a computational-efficiency perspective to the performance results in Tables S4 and S5. Image-only baselines achieve the highest raw throughput, but their weaker discrimination shows that speed alone does not characterize overall model performance. Within VLMs, PLIP and CLIP generally run faster, whereas CONCH provides stronger high-shot discrimination at a higher inference-time cost. Adaptation effects on throughput are method- and cohort-dependent: Classifier tuning is often the fastest VLM adaptation strategy, but its weaker robustness offsets this computational advantage; Adapter tuning offers a potentially useful trade-off in some settings, although its throughput advantage over Vanilla is not consistent across cohorts. For CONCH, Adapter improves throughput relative to Vanilla on Vanderbilt but not on Cornell. For future translational evaluation, model comparison should jointly consider discrimination, calibration, and inference throughput rather than any single metric; these benchmark-level comparisons do not by themselves establish clinical utility.

Taken together, within this retrospective two-cohort benchmark, the findings support a research-oriented model-selection pattern rather than a deployment rule. Pathology-specialized backbones provided stronger low-shot initializa-

tion, CONCH with Vanilla adaptation achieved the strongest discrimination when sufficient supervision was available, and Adapter tuning may warrant consideration when throughput-performance trade-offs are important. Calibration and margin analyses should accompany discrimination metrics in subsequent validation studies, particularly because confident errors persisted under cross-center variation. These observations can inform the design of prospective evaluations but should not be interpreted as recommendations for clinical deployment or safety.

5.4 Limitations

This study has several limitations. First, we use a single fixed text template for class prompts across all experiments. This design improves experimental control and comparability across backbones and adaptation strategies, but it also limits the richness of textual supervision and may underestimate the contribution of prompt design and textual priors to downstream vision-language performance. Future work should evaluate richer textual inputs, including morphology-specific descriptions, clinically relevant contextual information, and more structured reasoning-oriented prompts, together with systematic prompt sensitivity analyses.

Second, our few-shot experiments use repeated random class-balanced subsets while keeping the WSI-level train, validation, and test split fixed. This design was chosen to enable controlled shot-scaling comparisons, so that differences between 1-, 2-, 4-, and higher-shot settings primarily reflect the addition of training examples rather than changes in WSI composition, staining, scanner characteristics, tissue quality, or validation-set difficulty. The fixed validation WSIs also provide a consistent basis for early stopping and hyperparameter selection across shot settings and adaptation strategies. However, this design does not fully quantify split-to-split variability, and the conclusions may still be influenced by the particular WSI partition used in this study. Repeated WSI-level resampling or cross-validation would address this complementary question more directly and should be considered in future larger-scale evaluations.

Third, we formulate glomerular morphology as a single-label five-class task. Although this setting provides a controlled benchmark, it may not fully capture the full spectrum of morphologic variation observed in practice, particularly along transitions from ischemic change to segmental sclerosis and then global sclerosis. A more clinically faithful extension would represent lesions with multi-label annotations or lesion tags, allowing overlapping findings to be modeled explicitly while improving interpretability and better reflecting pathological continuity.

Fourth, this study is retrospective and includes a limited number of cohorts and centers. Although the Vanderbilt

cohort provides a cross-center evaluation under a distinct label schema and appearance distribution, the present design does not assess prospective workflow performance, human-AI interaction, or clinical safety, and it cannot establish broad generalizability across institutions, scanners, staining protocols, or patient populations. Accordingly, the throughput, calibration, and representation analyses should be interpreted as comparative research evidence rather than evidence of clinical readiness.

Finally, the expert label audit was limited to the Cornell subset. Similar annotation ambiguity may also affect Vanderbilt, especially near boundaries among disappearing, normal, obsolescent, solidified, and non-glomerular categories. However, Vanderbilt uses a different label schema from Cornell, and an equivalent blinded multi-rater re-annotation workflow was not available for the current study. Future work should extend expert-audit analysis to external cohorts using harmonized review protocols.

6. Conclusion

In this study, we frame fine-grained glomerular subtyping as a clinically motivated few-shot problem and use it to jointly examine supervision, adaptation strategy, and multimodal representation. We show that performance improves consistently with additional supervision, with meaningful gains already emerging from only a handful of labeled exemplars per class. Pathology-specialized VLM backbones offer a clear advantage in the extreme low-shot regime, but this gap narrows as more supervision becomes available. We also find that adaptation strategy governs not only accuracy but also how image and text features are organized. These observations indicate that multimodal alignment alone is not sufficient; how the model reshapes decision boundaries under limited labels is equally critical. Taken together, within this retrospective two-cohort setting, our results provide comparative evidence for model selection and annotation allocation under the evaluated data constraints. Alignment- and separability-based metrics may serve as candidate research indicators for future prospective multi-center validation, but they do not by themselves establish clinical readiness.

Acknowledgments

This research was supported by the WCM Radiology AIMI Fellowship, WCM CTSC 2027 Pilot Award, NIH 1R01HL174863-01A1(Sabuncu), and 1U54DK144866-01(Sabuncu).

Ethical Standards

The work follows appropriate ethical standards in conducting

research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

Data availability

In accordance with institutional regulations and in order to safeguard patient privacy and confidentiality, the datasets analyzed during this study cannot be shared and are not publicly available.

Appendix A. Detailed Mathematical Definitions

This appendix provides the detailed standard formulations underlying the compact final definitions retained in the main text.

A.1 Training Objectives

For a paired image x_i and class prompt y_i , let f_θ and g_ϕ denote the image and text encoders, respectively, and let τ denote the temperature. The normalized embeddings and temperature-scaled similarities are

$$\mathbf{v}_i = \frac{f_\theta(x_i)}{\|f_\theta(x_i)\|}, \quad \mathbf{t}_i = \frac{g_\phi(y_i)}{\|g_\phi(y_i)\|}, \quad s_{ij} = \frac{\mathbf{v}_i^\top \mathbf{t}_j}{\tau}. \quad (\text{S1})$$

The two directional losses and their symmetric average are

$$\begin{aligned} \mathcal{L}_{i \rightarrow t} &= -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s_{ii})}{\sum_{j=1}^N \exp(s_{ij})}, \\ \mathcal{L}_{t \rightarrow i} &= -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s_{ii})}{\sum_{j=1}^N \exp(s_{ji})}, \\ \mathcal{L}_{\text{ITC}} &= \frac{1}{2} (\mathcal{L}_{i \rightarrow t} + \mathcal{L}_{t \rightarrow i}). \end{aligned} \quad (\text{S2})$$

For Classifier tuning, $z_{i,c}$ denotes the logit for class c among C classes, and the multiclass cross-entropy objective is

$$p_{i,c} = \frac{\exp(z_{i,c})}{\sum_{r=1}^C \exp(z_{i,r})}, \quad \mathcal{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^N \log p_{i,y_i}. \quad (\text{S3})$$

A.2 Standard Evaluation Metrics

For the Silhouette score computed in the original embedding space with cosine distance, let n_c denote the number of samples in class c . The mean within-class distance a_i and

the minimum mean distance to another class b_i are

$$\begin{aligned} a_i &= \frac{1}{n_{y_i} - 1} \sum_{\substack{j \neq i \\ y_j = y_i}} d_{\cos}(\mathbf{x}_i, \mathbf{x}_j), \\ b_i &= \min_{c \neq y_i} \frac{1}{n_c} \sum_{y_j = c} d_{\cos}(\mathbf{x}_i, \mathbf{x}_j). \end{aligned} \quad (\text{S4})$$

The individual coefficient is $s_i = (b_i - a_i) / \max\{a_i, b_i\}$, and the dataset-level score reported in the main text is the mean of s_i over all samples.

For top-1 ECE, let $\hat{y}_i = \arg \max_c p_{i,c}$ and $q_i = \max_c p_{i,c}$ denote the predicted class and confidence. For each of the 15 equal-width bins B_m over $[0, 1]$, bin accuracy and confidence are

$$\begin{aligned} \text{acc}(B_m) &= \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}\{\hat{y}_i = y_i\}, \\ \text{conf}(B_m) &= \frac{1}{|B_m|} \sum_{i \in B_m} q_i. \end{aligned} \quad (\text{S5})$$

Empty bins contribute zero to the weighted ECE sum.

For Fleiss' κ , consider N samples rated by n raters into k classes, and let n_{ij} be the number of raters who assign sample i to class j . The per-sample and mean observed agreement are

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^k n_{ij}(n_{ij} - 1), \quad \bar{P} = \frac{1}{N} \sum_{i=1}^N P_i. \quad (\text{S6})$$

The marginal class proportion and chance-expected agreement are

$$p_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij}, \quad \bar{P}_e = \sum_{j=1}^k p_j^2. \quad (\text{S7})$$

Substitution of $\bar{P}$ and $\bar{P}_e$ into Eq. 8 yields the reported agreement statistic.

Appendix B. Additional Results on Vanderbilt Dataset

This appendix provides supplementary results on the Vanderbilt dataset to support the main findings and improve reproducibility.

B.1 Result Analysis on Vanderbilt

Fine-tuning strategy. Across Vanderbilt shot settings, adaptation strategy is the dominant source of performance variation (Table S1; Fig. S1). Vanilla fine-tuning provides the most consistent gains from low to high supervision, with CONCH with Vanilla reaching the strongest overall high-shot endpoint. Adapter and LoRA are generally intermediate and remain competitive in specific regimes, but

ROC Comparison of Backbones and Fine-tuning Strategies Across Shots

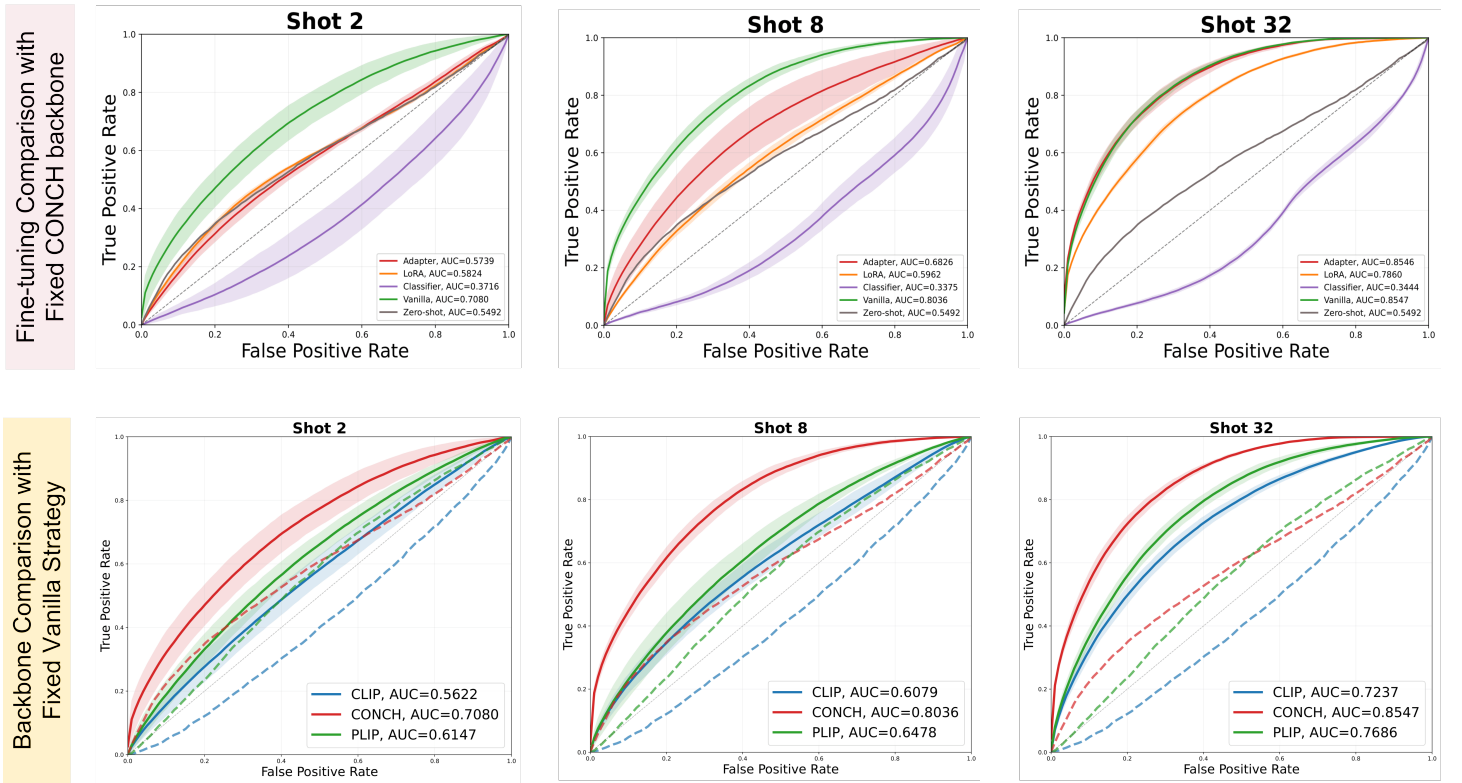


Figure S1: **ROC curves across models and fine-tuning methods on Vanderbilt dataset.** We compare backbones under fixed strategies and strategies under fixed backbones by plotting TPR versus FPR across increasing shot counts.

their trajectories are less uniformly strong across all metrics. By contrast, Classifier tuning is the least stable strategy on Vanderbilt and frequently degrades as shot increases, including severe collapse for CONCH and CLIP in both Accuracy and F1.

Backbone domain knowledge. Backbone pretraining domain relevance is clearly reflected in both low-shot and high-shot behavior. In 0-shot AUC, pathology-specialized backbones substantially exceed general CLIP, indicating a stronger pathology prior before adaptation. With additional supervision, CONCH preserves the highest ceiling across metrics, while PLIP remains competitive but typically below CONCH at medium-to-high shots. CLIP improves with supervision, especially under Vanilla/LoRA, but remains behind pathology-native backbones in the strongest Vanderbilt settings. This pattern is consistent with a domain-knowledge effect that improves both initialization quality and attainable high-shot performance.

Cross-modal indicators. Cross-modal trajectories follow the expected direction with increasing supervision (Fig. S3): Alignment tends to rise, and Similarity Gap moves from near-zero toward positive values. UMAP/KDE visualizations (Fig. S4) and geometric summaries (Table S3) show

progressively clearer organization at higher shots, especially for CONCH with Vanilla and CONCH with Adapter. However, compared with the Cornell trend strength, Vanderbilt shows flatter curves and more method overlap in several settings, indicating weaker separation in representation quality despite directional improvement.

Calibration and boundary margin. Confidence-related behavior is more heterogeneous than discrimination metrics on Vanderbilt (Table S2). Even when AUC and Accuracy improve, ECE often remains high at medium and high shots, and Classifier tuning shows pronounced calibration failure in high-shot settings. Boundary margin generally increases with supervision for stronger methods, but high margin does not always correspond to high correctness. This mismatch indicates persistent confident-error risk under Vanderbilt domain conditions.

B.2 Cross-Center Comparison

Comparing Cornell and Vanderbilt reveals shared directional trends but different effect magnitudes. In both centers, performance improves with shot count, pathology-specialized backbones outperform general CLIP, and Vanilla is usually the most reliable adaptation strategy at medium-to-high

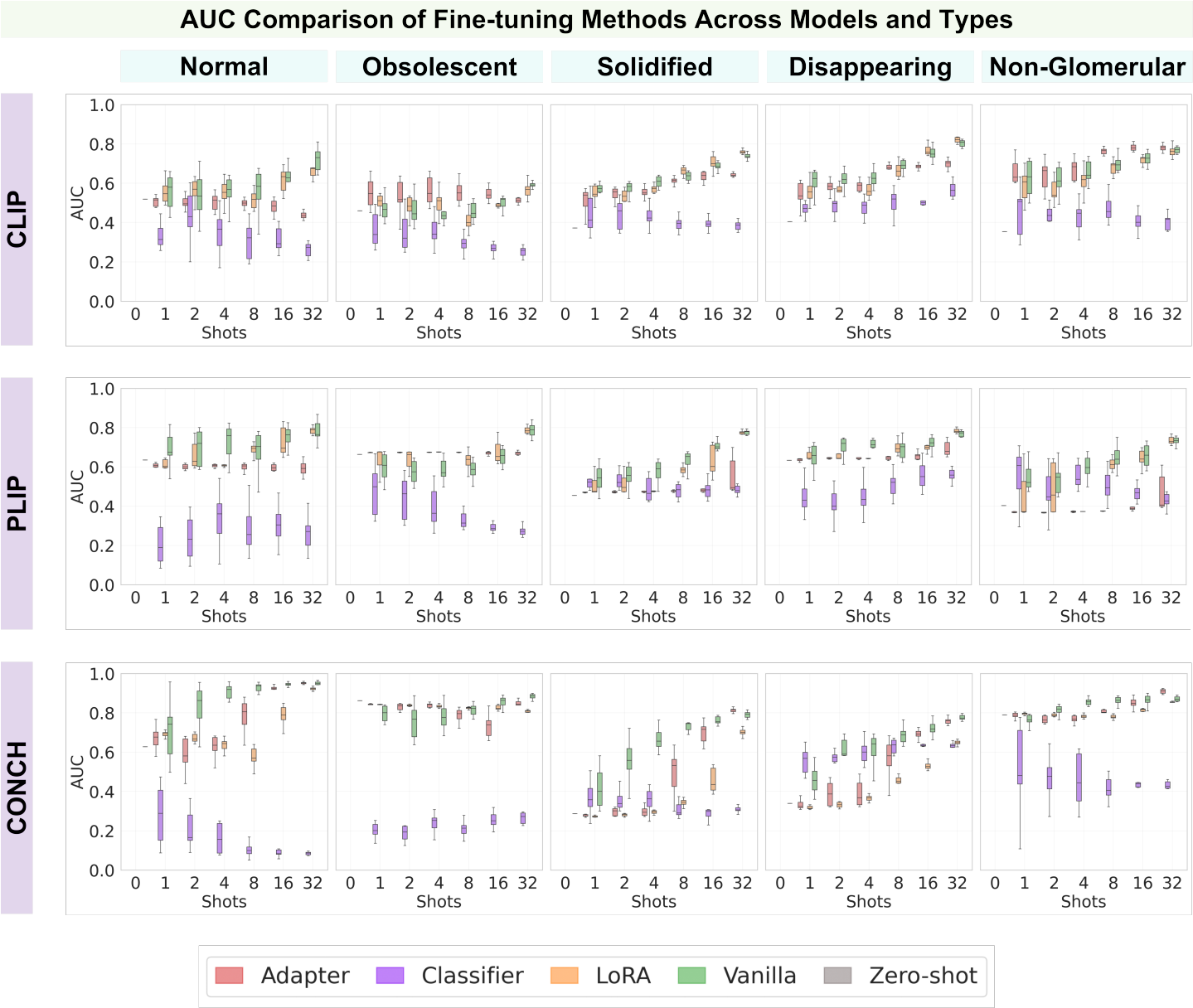


Figure S2: **Discrimination across subtypes on Vanderbilt.** This figure summarizes per-class discrimination performance across 10 runs. For each glomerular subtype, we report the distribution of AUC across VLM backbones, fine-tuning strategies, and shot settings.

shots. However, Vanderbilt exhibits stronger method sensitivity and weaker absolute discrimination endpoints, especially in Accuracy and F1, while Cornell shows steeper and more stable gains across multiple backbone-strategy pairs.

The largest center-level divergence is Classifier tuning. On Cornell, Classifier tuning can remain competitive for CONCH (including strong AUC at higher shots), whereas on Vanderbilt it degrades substantially across backbones and metrics, with poor robustness as supervision grows. A second divergence appears in confidence quality: Cornell high-shot settings more often achieve jointly favorable discrimination and calibration, while Vanderbilt retains higher ECE in many high-performing configurations. Taken together, these results support a consistent cross-center

conclusion: trend direction is reproducible, but practical reliability remains center-dependent and requires strategy-specific validation under external distribution shift.

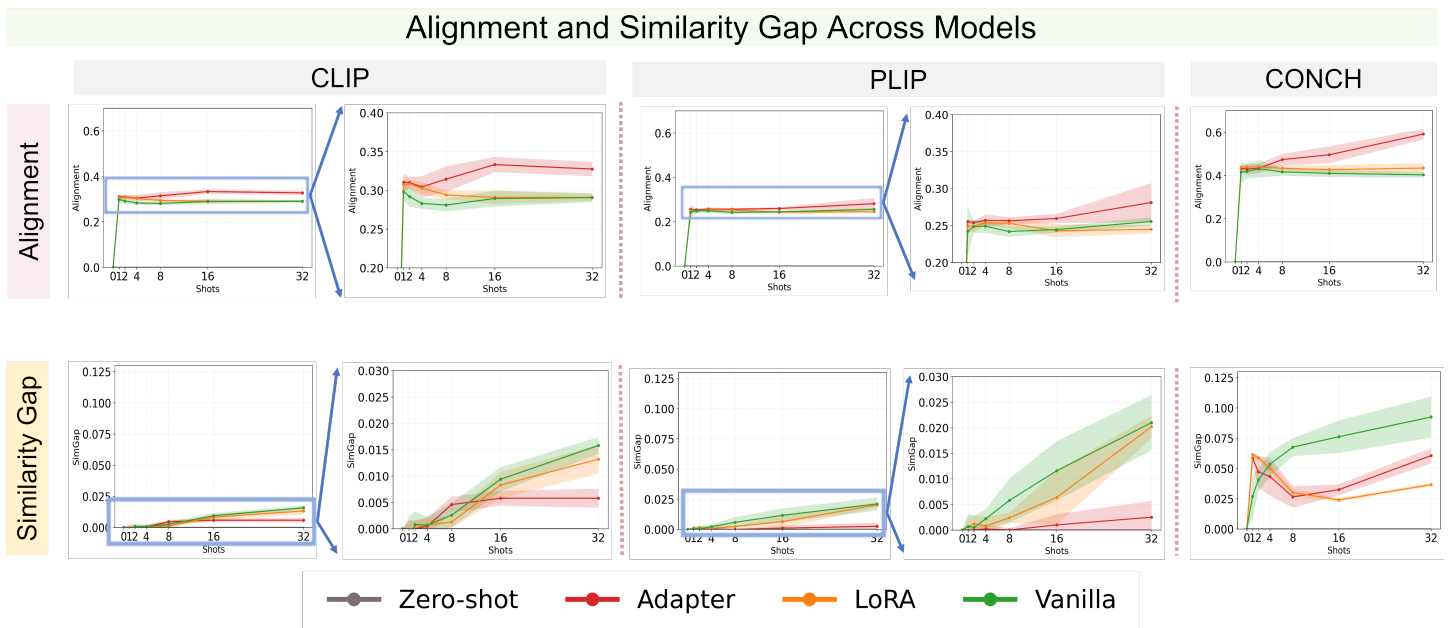


Figure S3: Alignment and similarity gap across models on Vanderbilt. This figure shows shot-wise trajectories of Alignment and Similarity Gap for CLIP, PLIP, and CONCH under different fine-tuning strategies. Similarity Gap is defined as mean positive class-prompt similarity minus mean negative class-prompt similarity. CLIP and PLIP are shown in both full-scale and zoomed-in panels for readability.

KDE and UMAP Visualization of the CONCH Model Under Various Fine-Tuning Methods and Shot Settings

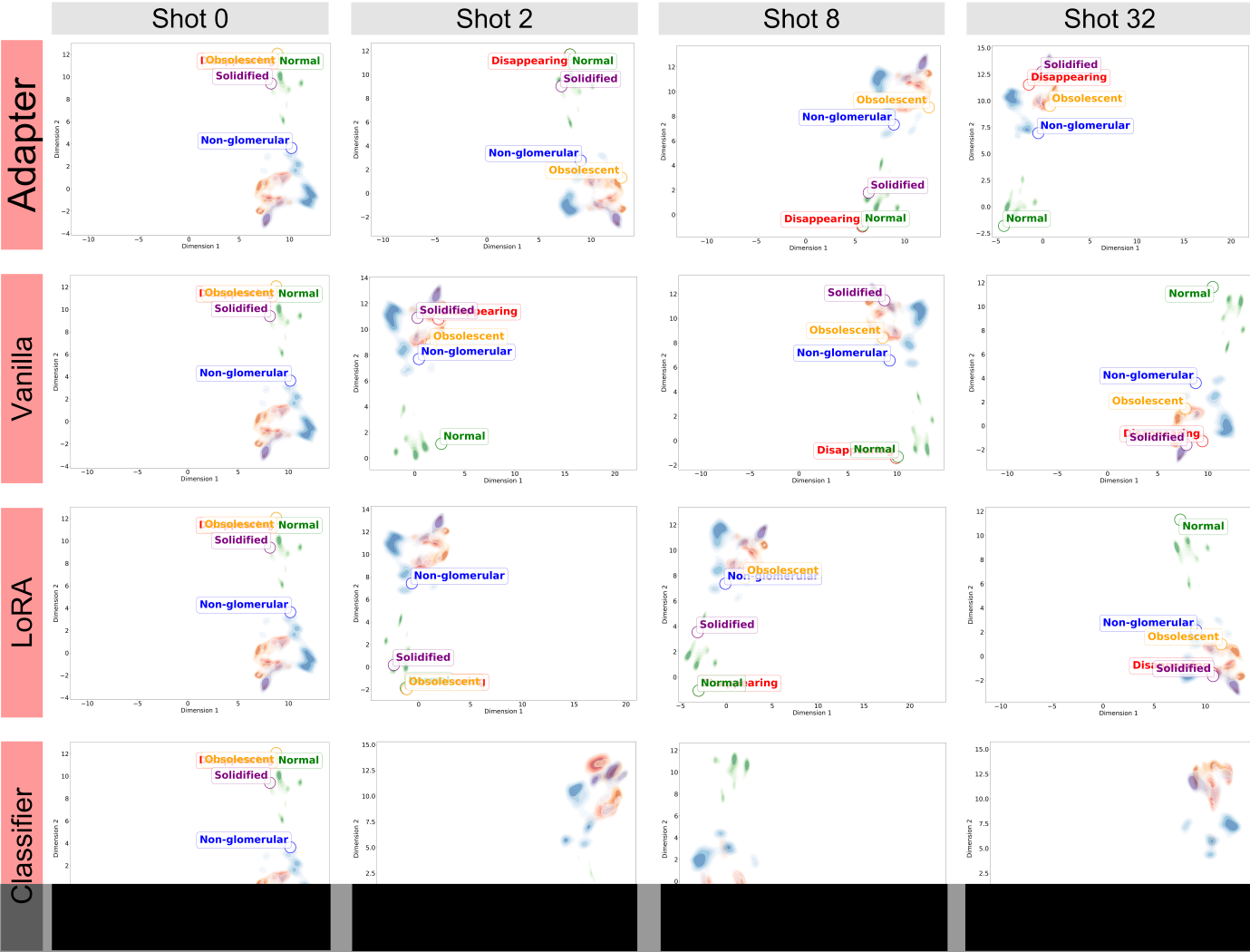


Figure S4: **UMAP/KDE visualization of embeddings across shots on Vanderbilt.** This figure shows text and image embeddings from the best CONCH run across shot settings. For Classifier tuning, text is not updated and only image-side embeddings are shown. Specifically, for Classifier tuning we visualize the penultimate hidden representation of the tuned MLP head (fed by fixed 512-D CONCH features), which changes as the head is updated. The companion Table S3 reports Intra-Class Distance and Silhouette Score.

Model	Method	Accuracy						
		0	1	2	4	8	16	32
CLIP Radford et al. (2021)	Adapter Gao et al. (2024)	0.2873±0.0000	0.1898±0.0688	0.2118±0.0615	0.2364±0.0630	0.3426±0.0316	0.3575±0.0182	0.3514±0.0174
	Classifier Ioffe and Szegedy (2015)	0.2873±0.0000	0.1456±0.0739	0.1091±0.0450	0.1079±0.0201	0.0935±0.0275	0.0796±0.0176	0.0691±0.0070
	LoRA Hu et al. (2022)	0.2873±0.0000	0.2531±0.0667	0.2683±0.0577	0.2820±0.0703	0.2667±0.0400	0.3407±0.0364	0.4025±0.0283
	Vanilla Srinivasan et al. (2024)	0.2873±0.0000	0.2505±0.0851	0.2924±0.0668	0.2699±0.0658	0.3245±0.0460	0.3266±0.0404	0.4023±0.0267
CONCH Lu et al. (2024)	Adapter Gao et al. (2024)	0.3281 ±0.0000	0.3576±0.0386	0.2977±0.0677	0.3168±0.0518	0.3639±0.0520	0.4173±0.0291	0.5053±0.0321
	Classifier Ioffe and Szegedy (2015)	0.3281 ±0.0000	0.1283±0.0608	0.0812±0.0380	0.1461±0.1012	0.0574±0.0251	0.0504±0.0112	0.0475±0.0088
	LoRA Hu et al. (2022)	0.3281 ±0.0000	0.3666 ±0.0065	0.3657±0.0114	0.3421±0.0274	0.2944±0.0317	0.3515±0.0297	0.4421±0.0067
	Vanilla Srinivasan et al. (2024)	0.3281 ±0.0000	0.2856±0.0722	0.3867 ±0.0712	0.4416 ±0.0542	0.4689 ±0.0267	0.4623 ±0.0301	0.5111 ±0.0441
PLIP Huang et al. (2023)	Adapter Gao et al. (2024)	0.1773±0.0000	0.1691±0.0139	0.1681±0.0278	0.1776±0.0352	0.1661±0.0127	0.1944±0.0455	0.2124±0.0331
	Classifier Ioffe and Szegedy (2015)	0.1773±0.0000	0.1414±0.0560	0.1326±0.0497	0.1324±0.0317	0.1063±0.0221	0.1067±0.0313	0.0989±0.0230
	LoRA Hu et al. (2022)	0.1773±0.0000	0.1981±0.0475	0.2380±0.0816	0.2080±0.0949	0.3262±0.0367	0.3227±0.0457	0.3992±0.0280
	Vanilla Srinivasan et al. (2024)	0.1773±0.0000	0.2314±0.0574	0.2403±0.0700	0.3096±0.0399	0.3301±0.0667	0.3289±0.0478	0.3853±0.0243
ResNet-18 He et al. (2016)	Supervised Learning	0.1560±0.0000	0.3486±0.0002	0.2481±0.0006	0.2196±0.0772	0.1355±0.0815	0.2261±0.0824	0.2192±0.0641
SSL Ciga et al. (2022)	Supervised Learning	0.2504±0.0000	0.2529±0.0014	0.0650±0.0195	0.2145±0.0749	0.2430±0.0880	0.2072±0.0960	0.1568±0.0900

Model	Method	AUC						
		0	1	2	4	8	16	32
CLIP Radford et al. (2021)	Adapter Gao et al. (2024)	0.4096±0.0000	0.5454±0.0452	0.5519±0.0496	0.5658±0.0458	0.6215±0.0215	0.6243±0.0154	0.6149±0.0069
	Classifier Ioffe and Szegedy (2015)	0.4096±0.0000	0.4145±0.0594	0.4289±0.0407	0.4094±0.0407	0.3914±0.0275	0.3796±0.0335	0.3759±0.0161
	LoRA Hu et al. (2022)	0.4096±0.0000	0.5526±0.0285	0.5374±0.0340	0.5602±0.0291	0.5855±0.0258	0.6607±0.0283	0.7168±0.0189
	Vanilla Srinivasan et al. (2024)	0.4096±0.0000	0.5630±0.0349	0.5622±0.0335	0.5769±0.0278	0.6079±0.0280	0.6581±0.0280	0.7237±0.0139
CONCH Lu et al. (2024)	Adapter Gao et al. (2024)	0.5492±0.0000	0.5815±0.0070	0.5739±0.0081	0.5875±0.0225	0.6826±0.0596	0.7797±0.0186	0.8546±0.0108
	Classifier Ioffe and Szegedy (2015)	0.5492±0.0000	0.3852±0.0722	0.3716±0.0573	0.3818±0.0639	0.3375±0.0316	0.3393±0.0179	0.3444±0.0072
	LoRA Hu et al. (2022)	0.5492±0.0000	0.5840±0.0016	0.5824±0.0028	0.5806±0.0063	0.5962±0.0106	0.6828±0.0146	0.7860±0.0039
	Vanilla Srinivasan et al. (2024)	0.5492±0.0000	0.6252 ±0.0000	0.7080 ±0.0441	0.7560 ±0.0294	0.8036 ±0.0148	0.8289 ±0.0175	0.8547 ±0.0085
PLIP Huang et al. (2023)	Adapter Gao et al. (2024)	0.5502±0.0000	0.5542±0.0101	0.5577±0.0157	0.5626±0.0220	0.5579±0.0104	0.5674±0.0193	0.5928±0.0476
	Classifier Ioffe and Szegedy (2015)	0.5502±0.0000	0.4485±0.0649	0.4156±0.0367	0.4374±0.0462	0.4199±0.0337	0.4201±0.0269	0.4043±0.0215
	LoRA Hu et al. (2022)	0.5502±0.0000	0.5749±0.0330	0.5905±0.0389	0.5781±0.0474	0.6412±0.0177	0.6766±0.0464	0.7719±0.0095
	Vanilla Srinivasan et al. (2024)	0.5502±0.0000	0.6028±0.0264	0.6147±0.0287	0.6370±0.0269	0.6478±0.0345	0.6962±0.0233	0.7686±0.0146
ResNet-18 He et al. (2016)	Supervised Learning	0.5065±0.0000	0.5681±0.0002	0.4338±0.0007	0.5167±0.0544	0.4761±0.0364	0.5245±0.0506	0.5262±0.0605
SSL Ciga et al. (2022)	Supervised Learning	0.5588 ±0.0000	0.5566±0.0005	0.4917±0.0209	0.4872±0.0385	0.4971±0.0677	0.5346±0.0645	0.4524±0.0588

Model	Method	F1						
		0	1	2	4	8	16	32
CLIP Radford et al. (2021)	Adapter Gao et al. (2024)	0.1049±0.0000	0.1362±0.0438	0.1519±0.0314	0.1666±0.0353	0.2392±0.0159	0.2468±0.0101	0.2392±0.0099
	Classifier Ioffe and Szegedy (2015)	0.1049±0.0000	0.0987±0.0420	0.0774±0.0318	0.0894±0.0171	0.0801±0.0179	0.0711±0.0149	0.0620±0.0052
	LoRA Hu et al. (2022)	0.1049±0.0000	0.1796±0.0426	0.1879±0.0265	0.2026±0.0322	0.2182±0.0271	0.2676±0.0238	0.3131±0.0212
	Vanilla Srinivasan et al. (2024)	0.1049±0.0000	0.1706±0.0462	0.1878±0.0372	0.2132±0.0405	0.2528±0.0270	0.2608±0.0363	0.3096±0.0223
CONCH Lu et al. (2024)	Adapter Gao et al. (2024)	0.1894 ±0.0000	0.2331 ±0.0202	0.2006±0.0396	0.2253±0.0272	0.2762±0.0394	0.3351±0.0224	0.4111±0.0240
	Classifier Ioffe and Szegedy (2015)	0.1894 ±0.0000	0.0833±0.0332	0.0602±0.0270	0.0852±0.0357	0.0547±0.0161	0.0495±0.0117	0.0492±0.0068
	LoRA Hu et al. (2022)	0.1894 ±0.0000	0.2298±0.0079	0.2417±0.0103	0.2416±0.0121	0.2197±0.0184	0.2783±0.0187	0.3463±0.0067
	Vanilla Srinivasan et al. (2024)	0.1894 ±0.0000	0.2242±0.0450	0.3017 ±0.0556	0.3552 ±0.0469	0.3795 ±0.0223	0.3750 ±0.0180	0.4118 ±0.0315
PLIP Huang et al. (2023)	Adapter Gao et al. (2024)	0.1361±0.0000	0.1369±0.0091	0.1377±0.0216	0.1453±0.0239	0.1391±0.0113	0.1548±0.0258	0.1749±0.0222
	Classifier Ioffe and Szegedy (2015)	0.1361±0.0000	0.1136±0.0376	0.0988±0.0207	0.1080±0.0182	0.0918±0.0150	0.0928±0.0226	0.0815±0.0099
	LoRA Hu et al. (2022)	0.1361±0.0000	0.1592±0.0326	0.1899±0.0638	0.1678±0.0726	0.2628±0.0249	0.2698±0.0325	0.3179±0.0224
	Vanilla Srinivasan et al. (2024)	0.1361±0.0000	0.1792±0.0403	0.1973±0.0472	0.2305±0.0291	0.2469±0.0508	0.2702±0.0313	0.3013±0.0195
ResNet-18 He et al. (2016)	Supervised Learning	0.0792±0.0000	0.1686±0.0001	0.0850±0.0001	0.1129±0.0310	0.0924±0.0433	0.1335±0.0413	0.1696±0.0428
SSL Ciga et al. (2022)	Supervised Learning	0.1205±0.0000	0.1234±0.0014	0.0386±0.0133	0.1262±0.0230	0.1528±0.0415	0.1206±0.0417	0.1025±0.0440

Table S1: **Few-shot classification performance on Vanderbilt.** Mean±SD Accuracy, AUC, and F1 across shot settings {0, 1, 2, 4, 8, 16, 32} for all backbone and adaptation pairs. Here, k -shot denotes k labeled images per class used for fine-tuning, and 0-shot denotes the zero-shot baseline. Bold indicates the best value within each shot column (higher is better).

Model	Method	ECE (Top-1)						
		0	1	2	4	8	16	32
CLIP Radford et al. (2021)	Adapter Gao et al. (2024)	0.0870±0.0000	0.1341±0.0581	0.1248±0.0411	0.1073±0.0427	0.0952±0.0318	0.1410±0.0332	0.1527±0.0423
	Classifier Ioffe and Szegedy (2015)	0.0870±0.0000	0.2010±0.0774	0.2289±0.0972	0.3632±0.1739	0.4579±0.1404	0.5120±0.1267	0.5888±0.0369
	LoRA Hu et al. (2022)	0.0870±0.0000	0.0841±0.0443	0.0610 ±0.0348	0.0577 ±0.0339	0.1809±0.0313	0.2759±0.0410	0.2606±0.0323
	Vanilla Srinivasan et al. (2024)	0.0870±0.0000	0.1920±0.1212	0.1451±0.0759	0.1143±0.0501	0.1928±0.1120	0.3420±0.0484	0.3088±0.0394
CONCH Lu et al. (2024)	Adapter Gao et al. (2024)	0.1447±0.0000	0.5159±0.0422	0.5653±0.0639	0.5235±0.0452	0.4120±0.0726	0.3881±0.0330	0.3368±0.0371
	Classifier Ioffe and Szegedy (2015)	0.1447±0.0000	0.0958±0.0510	0.1492±0.0247	0.1432±0.0943	0.3790±0.1544	0.5958±0.1671	0.7206±0.0133
	LoRA Hu et al. (2022)	0.1447±0.0000	0.5180±0.0112	0.4996±0.0168	0.5042±0.0269	0.5205±0.0365	0.4185±0.0263	0.3632±0.0060
	Vanilla Srinivasan et al. (2024)	0.1447±0.0000	0.4552±0.0770	0.3772±0.0707	0.3467±0.0590	0.3481±0.0320	0.3570±0.0354	0.3254±0.0413
PLIP Huang et al. (2023)	Adapter Gao et al. (2024)	0.0429±0.0000	0.2369±0.0313	0.2353±0.0539	0.2254±0.0434	0.2199±0.0338	0.1949±0.0752	0.2247±0.1242
	Classifier Ioffe and Szegedy (2015)	0.0429±0.0000	0.2584±0.1010	0.2107±0.1731	0.3677±0.1989	0.5620±0.0383	0.5245±0.0923	0.5485±0.0280
	LoRA Hu et al. (2022)	0.0429±0.0000	0.2303±0.0547	0.1777±0.0853	0.2135±0.0813	0.0482 ±0.0290	0.1580±0.1389	0.3000±0.0354
	Vanilla Srinivasan et al. (2024)	0.0429±0.0000	0.2790±0.1210	0.1788±0.0723	0.1176±0.0618	0.1997±0.1122	0.3297±0.0766	0.3367±0.0341
ResNet-18 He et al. (2016)	Supervised Learning	0.2112±0.0000	0.0223 ±0.0008	0.1735±0.0022	0.1328±0.0599	0.1827±0.0833	0.1069±0.0597	0.0818 ±0.0326
SSL Ciga et al. (2022)	Supervised Learning	0.0243 ±0.0000	0.0295±0.0012	0.2131±0.0274	0.0757±0.0155	0.0818±0.0348	0.0847 ±0.0549	0.1102±0.0581

Model	Method	Boundary Margin (Mean)						
		0	1	2	4	8	16	32
CLIP Radford et al. (2021)	Adapter Gao et al. (2024)	0.0013±0.0000	0.0807±0.0361	0.0901±0.0316	0.0915±0.0311	0.2123±0.0647	0.2749±0.0407	0.2781±0.0606
	Classifier Ioffe and Szegedy (2015)	0.0013±0.0000	0.1139±0.0468	0.1166±0.0966	0.2580±0.1813	0.3326±0.1363	0.3833±0.1299	0.4605±0.0483
	LoRA Hu et al. (2022)	0.0013±0.0000	0.0844±0.0236	0.0793±0.0233	0.0787±0.0298	0.1989±0.0323	0.3947±0.0476	0.4592±0.0379
	Vanilla Srinivasan et al. (2024)	0.0013±0.0000	0.1888±0.0642	0.1909±0.0614	0.1366±0.0300	0.2732±0.0842	0.4595±0.0205	0.5216±0.0277
CONCH Lu et al. (2024)	Adapter Gao et al. (2024)	0.0099±0.0000	0.7657±0.0305	0.7441±0.0368	0.7087±0.0649	0.6133±0.0537	0.6634±0.0224	0.7206 ±0.0184
	Classifier Ioffe and Szegedy (2015)	0.0099±0.0000	0.0128±0.0114	0.0231±0.0141	0.0426±0.0498	0.2173±0.1444	0.4645±0.1902	0.6049±0.0245
	LoRA Hu et al. (2022)	0.0099±0.0000	0.7878 ±0.0162	0.7512 ±0.0290	0.7145 ±0.0204	0.6587±0.0122	0.6024±0.0170	0.6656±0.0057
	Vanilla Srinivasan et al. (2024)	0.0099±0.0000	0.5539±0.0508	0.6001±0.0677	0.6389±0.0348	0.6846 ±0.0295	0.6863 ±0.0331	0.6957±0.0335
PLIP Huang et al. (2023)	Adapter Gao et al. (2024)	0.0012±0.0000	0.1698±0.0247	0.1668±0.0280	0.1673±0.0136	0.1491±0.0301	0.1526±0.0654	0.2038±0.1620
	Classifier Ioffe and Szegedy (2015)	0.0012±0.0000	0.1728±0.0720	0.1204±0.1606	0.2904±0.1981	0.4717±0.0547	0.4235±0.1052	0.4406±0.0488
	LoRA Hu et al. (2022)	0.0012±0.0000	0.1875±0.0342	0.1742±0.0234	0.1730±0.0181	0.1215±0.0151	0.2423±0.1584	0.4967±0.0216
	Vanilla Srinivasan et al. (2024)	0.0012±0.0000	0.2693±0.1311	0.1617±0.0329	0.1697±0.0732	0.2475±0.1235	0.4444±0.0975	0.5302±0.0482
ResNet-18 He et al. (2016)	Supervised Learning	0.0818 ±0.0000	0.0811±0.0005	0.1743±0.0030	0.0928±0.0309	0.0851±0.0308	0.0782±0.0176	0.0534±0.0070
SSL Ciga et al. (2022)	Supervised Learning	0.0286±0.0000	0.0275±0.0002	0.0669±0.0189	0.0261±0.0058	0.0216±0.0043	0.0276±0.0077	0.0254±0.0070

Table S2: **ECE and boundary margin on Vanderbilt.** ECE (Top-1) and boundary margin mean scores (Mean±SD) across shots {0, 1, 2, 4, 8, 16, 32}. Column-wise best values are bolded (lower is better for ECE; higher is better for boundary margin).

Model	Method	Intra-Class Distance							Silhouette Score						
		0	1	2	4	8	16	32	0	1	2	4	8	16	32
CLIP Radford et al. (2021)	Adapter Gao et al. (2024)	2.4820	2.4937	2.5021	2.3861	2.5315	2.7667	2.7161	-0.0291	-0.0342	-0.0334	-0.0379	-0.0365	-0.0326	-0.0364
	Classifier Ioffe and Szegedy (2015)	2.4820	3.1019	3.0256	2.7763	2.8256	2.8579	2.7962	-0.0291	0.0044	0.0468	0.0548	0.0087	0.0488	0.0484
	LoRA Hu et al. (2022)	2.4820	2.5154	2.5903	2.5607	2.5249	2.3678	2.3518	-0.0291	-0.0329	-0.0370	-0.0402	-0.0329	-0.0051	-0.0155
	Vanilla Srinivasan et al. (2024)	2.4820	2.6004	2.6157	2.5036	2.5971	2.5205	2.3991	-0.0291	-0.0336	-0.0312	-0.0332	-0.0176	0.0039	0.0300
CONCH Lu et al. (2024)	Adapter Gao et al. (2024)	2.3550	2.3776	2.3699	2.3708	2.3757	2.3952	2.3008	0.1032	0.1045	0.1047	0.0801	0.0998	0.0403	0.1108
	Classifier Ioffe and Szegedy (2015)	2.3550	2.8274	2.8359	2.7550	2.7543	2.7021	2.7898	0.1032	0.1917	0.1596	0.1622	0.1811	0.1547	0.1842
	LoRA Hu et al. (2022)	2.3550	2.3608	2.3951	2.3772	2.3838	2.3512	2.3034	0.1032	0.1026	0.1016	0.1007	0.0959	0.0933	0.0790
	Vanilla Srinivasan et al. (2024)	2.3550	2.3704	2.3716	2.3556	2.3633	2.3199	2.2636	0.1032	0.0982	0.0994	0.0972	0.0981	0.0978	0.1125
PLIP Huang et al. (2023)	Adapter Gao et al. (2024)	3.2031	3.1782	3.1748	3.2420	3.2093	3.2110	3.1746	0.0316	0.0315	0.0288	0.0314	0.0315	0.0314	0.0301
	Classifier Ioffe and Szegedy (2015)	3.2031	3.1665	3.2837	3.3406	3.2480	3.6128	3.2242	0.0316	0.0373	0.0566	0.0700	0.0724	0.0546	0.0740
	LoRA Hu et al. (2022)	3.2031	3.2644	3.1885	3.2145	3.1509	3.1498	2.9115	0.0316	0.0316	0.0322	0.0315	0.0321	0.0316	0.0625
	Vanilla Srinivasan et al. (2024)	3.2031	3.2055	3.2171	3.1875	3.2082	3.2044	3.0516	0.0316	0.0312	0.0302	0.0294	0.0439	0.0651	0.0837

Table S3: **Intra-Class Distance and Silhouette Score on Vanderbilt.** Few-shot geometric metrics under three backbones: CLIP, CONCH, and PLIP across shot settings {0, 1, 2, 4, 8, 16, 32} and four fine-tuning strategies. Intra-Class Distance measures within-class compactness in the UMAP space (lower is better), and Silhouette Score measures cluster separability in the original embedding space using cosine distance (higher is better). For Classifier tuning, both metrics are computed from the penultimate representation of the tuned MLP head.

Model	Method	Samples/s ($\mu \pm \sigma$)
CLIP Radford et al. (2021)	0-shot	38.8668 $\pm$ 0.0000
	Adapter Gao et al. (2024)	39.1536 $\pm$ 0.0325
	Classifier Ioffe and Szegedy (2015)	41.5737 $\pm$ 0.4636
	LoRA Hu et al. (2022)	41.2589 $\pm$ 0.4928
	Vanilla Srinivasan et al. (2024)	40.7104 $\pm$ 0.6712
CONCH Lu et al. (2024)	0-shot	29.3611 $\pm$ 0.0000
	Adapter Gao et al. (2024)	30.6423 $\pm$ 0.2150
	Classifier Ioffe and Szegedy (2015)	31.4967 $\pm$ 0.0799
	LoRA Hu et al. (2022)	29.5964 $\pm$ 0.2099
	Vanilla Srinivasan et al. (2024)	30.6389 $\pm$ 0.4046
PLIP Huang et al. (2023)	0-shot	42.3505 $\pm$ 0.0000
	Adapter Gao et al. (2024)	43.6870 $\pm$ 0.4841
	Classifier Ioffe and Szegedy (2015)	42.4355 $\pm$ 0.5230
	LoRA Hu et al. (2022)	43.4285 $\pm$ 0.7742
	Vanilla Srinivasan et al. (2024)	42.6651 $\pm$ 0.6393
ResNet-18 He et al. (2016)	Supervised Learning	47.0729 $\pm$ 1.2160
SSL Ciga et al. (2022)	Supervised Learning	52.5185 $\pm$ 3.5419

Table S4: **Inference throughput on Cornell.** Samples processed per second during inference on a 1,000-sample timing subset, summarized as mean $\pm$ SD. For VLMs, the 0-shot row reports base model inference timing, while Adapter, Classifier, LoRA, and Vanilla rows summarize shots {1, 2, 4, 8, 16, 32}. ResNet-18 and SSL rows summarize shots {0, 1, 2, 4, 8, 16, 32}. Higher values indicate faster inference; bold indicates the highest mean throughput.

Model	Method	Samples/s ($\mu \pm \sigma$)
CLIP Radford et al. (2021)	0-shot	32.6514 $\pm$ 0.0000
	Adapter Gao et al. (2024)	35.9904 $\pm$ 5.9719
	Classifier Ioffe and Szegedy (2015)	39.5126 $\pm$ 0.4868
	LoRA Hu et al. (2022)	38.7514 $\pm$ 0.5405
	Vanilla Srinivasan et al. (2024)	36.5264 $\pm$ 6.3853
CONCH Lu et al. (2024)	0-shot	27.6750 $\pm$ 0.0000
	Adapter Gao et al. (2024)	28.3857 $\pm$ 0.6497
	Classifier Ioffe and Szegedy (2015)	31.2421 $\pm$ 0.3824
	LoRA Hu et al. (2022)	29.0971 $\pm$ 0.6714
	Vanilla Srinivasan et al. (2024)	27.4787 $\pm$ 4.3375
PLIP Huang et al. (2023)	0-shot	39.7531 $\pm$ 0.0000
	Adapter Gao et al. (2024)	41.6119 $\pm$ 0.7270
	Classifier Ioffe and Szegedy (2015)	40.0198 $\pm$ 0.9968
	LoRA Hu et al. (2022)	40.1168 $\pm$ 1.6200
	Vanilla Srinivasan et al. (2024)	39.8413 $\pm$ 5.0300
ResNet-18 He et al. (2016)	Supervised Learning	43.1835 $\pm$ 0.9697
SSL Ciga et al. (2022)	Supervised Learning	45.8984 $\pm$ 1.8326

Table S5: **Inference throughput on Vanderbilt.** Samples processed per second during inference on a 1,000-sample timing subset, summarized as mean $\pm$ SD. For VLMs, the 0-shot row reports base model inference timing, while Adapter, Classifier, LoRA, and Vanilla rows summarize shots {1, 2, 4, 8, 16, 32}. ResNet-18 and SSL rows summarize shots {0, 1, 2, 4, 8, 16, 32}. Higher values indicate faster inference; bold indicates the highest mean throughput.

References

- Ingeborg M Bajema, Suzanne Wilhelmus, Charles E Alpers, Jan A Bruijn, Robert B Colvin, H Terence Cook, Vivette D D'Agati, Franco Ferrario, Mark Haas, J Charles Jennette, et al. Revision of the international society of nephrology/renal pathology society classification for lupus nephritis: clarification of definitions, and modified national institutes of health activity and chronicity indices. *Kidney International*, 93(4):789–796, 2018.
- Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Cetin, Eren Halici, Hunter Jackson, Richard Chen, Fabian Both, Jorg Franke, Heidi Kusters-Vandeveld, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, February 2019. .
- Mohsin Bilal, Aadam, Manahil Raza, Youssef Altherwy, Anas Alsuhaibani, Abdulrahman Abduljabbar, Fahdah Almarshad, Paul Golding, and Nasir Rajpoot. Foundation models in computational pathology: A review of challenges, opportunities, and impact, 2025. URL <https://arxiv.org/abs/2502.08333>.
- Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. ISSN 2078-2489. . URL <https://www.mdpi.com/2078-2489/11/2/125>.
- Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Mirafior, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.
- Liu Chan, Yang Danyi, and Chao Chen. A new index for the outcome of focal segmental glomerulosclerosis. *Scientific Reports*, 14(1):8278, 2024.
- Dibaloke Chanda, Milan Aryal, Nasim Yahya Soltani, and Masoud Ganji. A new era in computational pathology: A survey on foundation and vision-language models, 2024. URL <https://arxiv.org/abs/2408.14496>.
- Qi Chen, Ruoshan Zhao, Sinuo Wang, Vu Minh Hieu Phan, Anton van den Hengel, Johan Verjans, Zhibin Liao, Minh-Son To, Yong Xia, Jian Chen, et al. A survey of medical vision-and-language applications and their techniques, 2024. URL <https://arxiv.org/abs/2411.12195>.
- Ying Chen, Yiya Yang, Yumei Liang, Manting Liu, Wei Xiao, and Xiaofang Hu. Retrospective analysis of crescent score in clinical prognosis of iga nephropathy. *Open Medicine*, 17(1):205–215, 2022.
- Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.
- Rebecca S Crowley, Gregory J Naus, Jimmie Stewart III, and Charles P Friedman. Development of visual diagnostic expertise in pathology: an information-processing study. *Journal of the American Medical Informatics Association*, 10(1):39–51, 2003.
- Can Cui, Han Liu, Quan Liu, Ruining Deng, Zuhayr Asad, Yaohong Wang, Shilin Zhao, Haichun Yang, Bennett A Landman, and Yuankai Huo. Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 626–635. Springer, 2022.
- Can Cui, Ruining Deng, Junchao Guo, Quan Liu, Tianyuan Yao, Haichun Yang, and Yuankai Huo. Enhancing physician flexibility: Prompt-guided multi-class pathological

- segmentation for diverse outcomes. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–8. IEEE, 2024.
- Ruining Deng, Can Cui, Lucas W Remedios, Shunxing Bao, R Michael Womick, Sophie Chiron, Jia Li, Joseph T Roland, Ken S Lau, Qi Liu, et al. Cross-scale attention guided multi-instance learning for crohn’s disease diagnosis with pathological images. In *International Workshop on Multiscale Multimodal Medical Imaging*, pages 24–33. Springer, 2022.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL <https://arxiv.org/abs/2010.11929>.
- Savannah R Duenweg, Samuel A Bobholz, Allison K Lowman, Margaret A Stebbins, Aleksandra Winiarz, Biprojit Nath, Fitzgerald Kyereme, Kenneth A Iczkowski, and Peter S LaViolette. Whole slide imaging (wsi) scanner differences influence optical and computed properties of digitized prostate cancer histology. *Journal of Pathology Informatics*, 14:100321, 2023.
- Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, CAMELYON16 consortium, Meyke Hermsen, Quirine F Manson, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.
- Chunyue Feng and Fei Liu. Artificial intelligence in renal pathology: current status and future. *Biomolecules and Biomedicine*, 23(2):225, 2023.
- Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.
- 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.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. URL <https://proceedings.mlr.press/v70/guo17a.html>.
- Zhenhao Guo, Dingcheng Duan, Youwei Liang, Atharvraj Patil, and Pengtao Xie. Metabolitechat: A unified multimodal large language model for interactive metabolite analysis and functional insights. *bioRxiv*, pages 2025–11, 2025.
- Zhenhao Guo, Rachit Saluja, Tianyuan Yao, Quan Liu, Yuankai Huo, Benjamin Liechty, David J Pisapia, Kenji Ikemura, Mert R Sabuncu, Yihe Yang, et al. Glo-VLMs: Leveraging Vision-Language Models for Fine-Grained Diseased Glomerulus Classification. In *Medical Imaging 2026: Digital and Computational Pathology*, volume 13932, page 1393208, 2026.
- Matthew G Hanna, Liron Pantanowitz, Rajesh Dash, James H Harrison, Mustafa Deebajah, Joshua Pantanowitz, and Hooman H Rashidi. Future of artificial intelligence (ai)-machine learning (ml) trends in pathology and medicine. *Modern Pathology*, page 100705, 2025.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Mahdi S Hosseini, Babak Ehteshami Bejnordi, Vincent Quoc-Huy Trinh, Lyndon Chan, Danial Hasan, Xingwen Li, Stephen Yang, Taehyo Kim, Haochen Zhang, Theodore Wu, et al. Computational pathology: a survey review and the way forward. *Journal of Pathology Informatics*, 15:100357, 2024.
- Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 3–13. Springer, 2021.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2):3, 2022.
- 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.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

- Yihong Jin, Ze Yang, Juntian Liu, and Xinhe Xu. Anomaly detection and early warning mechanism for intelligent monitoring systems in multi-cloud environments based on LLM. In *2025 5th International Symposium on Computer Technology and Information Science (ISCTIS)*, pages 167–170, 2025. .
- Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022.
- Qunjuan Lei, Xiaoshuai Hou, Xumeng Liu, Dongmei Liang, Yun Fan, Feng Xu, Shaoshan Liang, Dandan Liang, Jing Yang, Guotong Xie, et al. Artificial intelligence assists identification and pathologic classification of glomerular lesions in patients with diabetic nephropathy. *Journal of Translational Medicine*, 22(1):397, 2024.
- Dong Li, Guihong Wan, Xintao Wu, Xinyu Wu, Ajit J Nirmal, Christine G Lian, Peter K Sorger, Yevgeniy R Semenov, and Chen Zhao. A survey on computational pathology foundation models: Datasets, adaptation strategies, and evaluation tasks, 2025a. URL <https://arxiv.org/abs/2501.15724>.
- Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey, 2024. URL <https://arxiv.org/abs/2411.17040>.
- Xueyuan Li, Can Cui, Ruining Deng, Yucheng Tang, Quan Liu, Tianyuan Yao, Shunxing Bao, Naweel Chowdhury, Haichun Yang, and Yuankai Huo. Fine-grained multiclass nuclei segmentation with molecular empowered all-in-sam model. *Journal of Medical Imaging*, 12(5):057501–057501, 2025b.
- Fang Lu, Shan Hou, Chang Liu, Qing Li, Lin Wu, Ying Pan, Yangfan Wu, Hua Shu, Bo Zhang, Huijuan Mao, et al. Clinical significance of focal segmental glomerulosclerosis subclassification in iga nephropathy. *Nephrology Dialysis Transplantation*, page gfaf162, 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.
- Yuzhe Lu, Haichun Yang, Zheyu Zhu, Ruining Deng, Agnes B Fogo, and Yuankai Huo. Improve global glomerulosclerosis classification with imbalanced data using CircleMix augmentation. In *Medical Imaging 2021: Image Processing*, volume 11596, page 1159609, 2021. .
- Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016.
- Sajid Nazir, Diane M Dickson, and Muhammad Usman Akram. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine*, 156:106668, 2023.
- Mieko Ochi, Daisuke Komura, and Shumpei Ishikawa. Pathology foundation models. *JMA Journal*, 8(1):121–130, 2025.
- Eva Pachetti and Sara Colantonio. A systematic review of few-shot learning in medical imaging. *Artificial Intelligence in Medicine*, 156:102949, 2024.
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, 2020. .
- 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.
- M Rafieian-Kopaei, A Baradaran, and H Nasri. Significance of extracapillary proliferation in iga-nephropathy patients with regard to clinical and histopathological variables. *Hippokratia*, 17(3):258, 2013.
- Christopher J Roth, David A Clunie, David J Vining, Seth J Berkowitz, Alejandro Berlin, Jean-Pierre Bissonnette, Shawn D Clark, Toby C Cornish, Monief Eid, Cree M Gaskin, et al. Multispecialty enterprise imaging workgroup consensus on interactive multimedia reporting current state and road to the future: Himss-siim collaborative white paper. *Journal of Digital Imaging*, 34(3):495–522, 2021.
- Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. .
- Peter Schnuelle. Renal biopsy for diagnosis in kidney disease: indication, technique, and safety. *Journal of Clinical Medicine*, 12(19):6424, 2023.
- Sanjeev Sethi and Fernando C Fervenza. Standardized classification and reporting of glomerulonephritis. *Nephrology Dialysis Transplantation*, 34(2):193–199, 2019.

- Fereshteh Shakeri, Malik Boudiaf, Sina Mohammadi, Ivaxi Sheth, Mohammad Havaei, Ismail Ben Ayed, and Samira Ebrahimi Kahou. Fhist: a benchmark for few-shot classification of histological images, 2022. URL <https://arxiv.org/abs/2206.00092>.
- Jiangbo Shi, Zeyu Gao, Haichuan Zhang, Pargorn Puttapirat, Chunbao Wang, Xiangrong Zhang, and Chen Li. Effects of annotation granularity in deep learning models for histopathological images. In *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 2702–2708. IEEE, 2019.
- Benjamin D Simon, Kutsev Bengisu Ozyoruk, David G Gelikman, Stephanie A Harmon, and Barış Türkbeş. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. *Diagnostic and Interventional Radiology*, 31(4): 303, 2025.
- Anabia Sohail, Iyyakutti Iyappan Ganapathi, Basit Alawode, Sajid Javed, Mohammed Bennamoun, and Arif Mahmood. Convlm: Context-guided vision-language model for fine-grained histopathology image classification. *Information Fusion*, page 103737, 2025.
- Krishna Prasad Varadarajan Srinivasan, Prasanth Gumpena, Madhusudhana Yattapu, and Vishal H Brahmbhatt. Comparative analysis of different efficient fine tuning methods of large language models (llms) in low-resource setting, 2024. URL <https://arxiv.org/abs/2405.13181>.
- Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Michael B Gotway, and Jianming Liang. Large-scale benchmarking and boosting transfer learning for medical image analysis. *Medical Image Analysis*, 102:103487, 2025.
- Thijs W Cohen Tervaert, Antien L Mooyaart, Kerstin Amann, Arthur H Cohen, H Terence Cook, Cinthia B Drachenberg, Franco Ferrario, Agnes B Fogo, Mark Haas, Emile De Heer, et al. Pathologic classification of diabetic nephropathy. *Journal of the American Society of Nephrology*, 21(4):556–563, 2010.
- Liang Wang and Luis Carvalho. Deviance matrix factorization. *Electronic Journal of Statistics*, 17(2):3762–3810, 2023.
- Liang Wang and Luis Carvalho. Multiclass roc, 2024. URL <https://arxiv.org/abs/2404.13147>.
- Liang Wang and Luis Carvalho. Computational approaches for exponential-family factor analysis. *Journal of Computational and Graphical Statistics*, pages 1–19, 2026.
- Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.
- Xiangxue Wang, Jingkai Zhang, Yuemei Xu, Yang Huang, Wenlong Ming, Yiping Jiao, Bicheng Liu, Xiangshan Fan, and Jun Xu. Glo-net: A dual task branch based neural network for multi-class glomeruli segmentation. *Computers in Biology and Medicine*, 186:109670, 2025. ISSN 0010-4825. . URL <https://www.sciencedirect.com/science/article/pii/S0010482525000204>.
- Cleo-Aron Weis, Jan Niklas Bindzus, Jonas Voigt, Marlen Runz, Svetlana Hertjens, Matthias M Gaida, Zoran V Popovic, and Stefan Porubsky. Assessment of glomerular morphological patterns by deep learning algorithms. *Journal of Nephrology*, 35(2):417–427, 2022.
- Parker C Wilson and Nidia Messias. How whole slide imaging and machine learning can partner with renal pathology, 2022.
- Working Group of the International IgA Nephropathy Network and the Renal Pathology Society, Daniel C. Cattran, Rosanna Coppo, H. Terence Cook, John Feehally, Ian S. D. Roberts, Stéphan Troyanov, Charles E. Alpers, Alessandro Amore, Jonathan Barratt, Francois Berthou, Stephen Bonsib, Jan A. Bruijn, Vivette D’Agati, Giuseppe D’Amico, Steven Emancipator, Francesco Emma, Franco Ferrario, Fernando C. Fervenza, Sandrine Florquin, Agnes Fogo, Colin C. Geddes, Hermann-Josef Groene, Mark Haas, Andrew M. Herzenberg, Prue A. Hill, Ronald J. Hogg, Stephen I. Hsu, J. Charles Jennette, Kensuke Joh, Bruce A. Julian, Tetsuya Kawamura, Fernand M. Lai, Chi Bon Leung, Lei-Shi Li, Philip K. T. Li, Zhi-Hong Liu, Bruce Mackinnon, Sergio Mezzano, F. Paolo Schena, Yasuhiko Tomino, Patrick D. Walker, Haiyan Wang, Jan J. Weening, Nori Yoshikawa, and Hong Zhang. The oxford classification of IgA nephropathy: Rationale, clinicopathological correlations, and classification. *Kidney International*, 76(5):534–545, September 2009. .
- Conghao Xiong, Hao Chen, and Joseph JY Sung. A survey of pathology foundation model: Progress and future directions, 2025. URL <https://arxiv.org/abs/2504.04045>.
- Xinhe Xu, Zhuoer Wang, Yihan Zhang, Yizhou Liu, Zhaoyue Wang, Zhihao Xu, Muhan Zhao, and Huaiying Luo. Style transfer: From stitching to neural networks. In *2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, pages 526–530. IEEE, 2024.

- Tianyuan Yao, Yuzhe Lu, Ruining Deng, Zheyu Zhu, Zuhayr Asad, Haichun Yang, Lee E Wheless, Agnes B Fogo, and Yuankai Huo. Self-supervised learning with large-scale web image mining for characterizing glomerular lesions. In *Medical Imaging 2022: Digital and Computational Pathology*, volume 12039, pages 160–166. SPIE, 2022.
- Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A Mazurowski, and Heung-II Suk. Domain generalization for medical image analysis: A review. *Proceedings of the IEEE*, 112(10):1583–1609, 2024.
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. URL <https://arxiv.org/abs/2205.01917>.
- Lining Yu, Mengmeng Yin, Ruining Deng, Quan Liu, Tianyuan Yao, Can Cui, Junlin Guo, Yu Wang, Yao-hong Wang, Shilin Zhao, et al. Glo-in-one-v2: holistic identification of glomerular cells, tissues, and lesions in human and mouse histopathology. *Journal of Medical Imaging*, 12(6):061406–061406, 2025.
- Sunyi Zheng, Xiaonan Cui, Yuxuan Sun, Jingxiong Li, Honglin Li, Yunlong Zhang, Pingyi Chen, Xueping Jing, Zhaoxiang Ye, and Lin Yang. Benchmarking pathclip for pathology image analysis. *Journal of Imaging Informatics in Medicine*, 38(1):422–438, 2025.