Beyond Sparse Captions: Aligning Slide-Level Text and Patch-Level Vision in Pathology

Daniel Shao[*]
MIT
MGB AI Institute
Harvard University
dshao@mit.edu

Dongmin Bang[*]
MGB AI Institute
Harvard University

Luca L. Weishaupt
MIT
MGB AI Institute
Harvard University

Sophia J. Wagner
MGB AI Institute
Harvard University

Nic G. Reitsam
University of Augsburg
MGB AI Institute
Harvard University

Ming Y. Lu
MGB AI Institute
Harvard University

Long P. Le
MGB AI Institute
Harvard University

Richard J. Chen[†]
MGB AI Institute
Harvard University

Faisal Mahmood[†]
MGB AI Institute
Harvard University

[*]Co-first authors
[†]Co-senior authors

Abstract

Contrastive vision-language (VL) pretraining in computational pathology has generally relied on curated image-caption datasets sourced from social media, educational videos, and research publications. These corpora are sparse relative to the diversity of tissue morphology and difficult to scale further, as pathologists do not routinely produce descriptive captions in clinical practice. We propose to instead leverage slide-level pathology reports which can be generated during routine diagnostic workflows in a scalable manner. A key challenge in this weakly-supervised setting is that pathology reports describe macroscopic, diagnostic findings while patch-level visual features describe microscopic findings. We introduce Locked Image Multiple Instance PreTraining (LIMIT), which addresses this gap by freezing a pretrained vision encoder and fine-tuning only the text encoder: report embeddings serve as cross-attention queries over bags of patch features, producing report-conditioned slide representations optimized via InfoNCE. Using LIMIT, we establish CONCH-Z, which aligns the text encoder of CONCH v1.5 to 335k clinical reports. Evaluated across 22 tasks spanning patch-level classification, tumor detection, and slide-level subtyping, CONCH-Z establishes state-of-the-art zero-shot performance over VL encoders trained on curated caption datasets, while simultaneously closing the gap with slide-level foundation models trained with substantially more parameters and compute. The code and model is available at https://github.com/mahmoodlab/LIMIT.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Foundation models have advanced computational pathology (CPath), enabling scalable analysis of digitized tissue sections for diagnosis, prognosis, and biomarker prediction [1, 2]. In particular, vision-language (VL) models have extended patch-level representation learning by aligning visual features with medical text, unlocking multimodal capabilities such as zero-shot classification, text-to-image retrieval, and slide-level inference without task-specific supervision [3–5]. VL encoders have consequently become integral to a growing number of CPath systems, serving as upstream modules for slide-level architectures [6, 7], explainability frameworks [8], and reasoning systems [9, 10].

Although the abundance of unlabeled tissue patches enables large-scale vision-only pretraining [11, 12], vision-language (VL) pretraining in computational pathology faces a fundamental data bottleneck. Because paired clinical data are rarely available in public repositories, existing pathology VL models rely on relatively small corpora compiled from social media platforms, educational videos, and open-access publications [13, 14, 3]. While these curated datasets have facilitated initial pathology VL encoders, they depend on descriptive patch captions written for academic or pedagogical purposes; in routine clinical workflows, pathologists instead author slide-level diagnostic reports that synthesize key findings across the specimen and naturally scale alongside clinical operations. Consequently, curated patch-level datasets (typically comprising around one million pairs) remain sparse relative to the vast morphological diversity of human tissue, exhibit a narrow linguistic scope, and cannot be easily expanded under current collection paradigms. Advancing VL alignment in pathology therefore requires moving beyond curated patch captions to directly leverage routine slide-level diagnostic reports.

This work addresses the challenge of VL alignment under weak supervision, where slide-level diagnostic text summarizes findings but only patch-level visual features are available. This setting mirrors the classical MIL formulation in CPath, which learns to aggregate patch representations for slide-level prediction without patch-level labels, but extended to the cross-modal case, where the supervisory signal is normalized, open-vocabulary diagnostic text rather than a fixed-vocabulary class label.

To this end, we propose **L**ocked **I**mage **M**ultiple **I**nstance Pre**T**raining (LIMIT), which freezes a pretrained vision encoder and fine-tunes only the text encoder by using report embeddings as cross-attention queries over bags of patch embeddings, producing report-conditioned slide representations that can be optimized via InfoNCE in slide classification pretext tasks. Using LIMIT, we present CONCH-Z, which aligns the text encoder of CONCHv1.5[3] to diagnostic text from 335,645 slide–text pairs for improved zero-shot cross-modal capabilities. Evaluated across diverse patch- and slide-level tasks, CONCH-Z reaches SOTA performance on zero-shot patch-level classification compared to VL encoders trained on curated caption datasets, while comparing favorably with slide-level foundation models (*e.g.*, PRISM, TITAN) that train dedicated slide representations (**Table 2**). Our main contributions are as follows:

- We formalize VL pretraining for pathology as a weakly-supervised cross-modal MIL problem and introduce LIMIT, a method that fine-tunes only the text encoder via report-guided cross-attention over frozen patch embeddings.
- We train CONCH-Z on normalized diagnostic text from 335,645 slide–text pairs and establish state-of-the-art zero-shot performance across diverse benchmarks, including pan-cancer subtyping, tumor detection, and patch-level classification.
- We show that slide-level diagnostic supervision improves patch- and slide-level visual-semantic alignment without local annotations or learned aggregation parameters, and validate LIMIT across additional pretrained VL models and a public training corpus.

2 Related Works

Vision-Language Models in Computational Pathology. Early patch-level VL models such as PLIP aligned pathology images with text from medical social media and academic sources [13], while CONCH and Quilt-1M scaled this paradigm to over one million curated image-caption pairs from PubMed Central, educational sources, YouTube, and social media [3, 14]. MUSK further expanded pretraining with 50 million pathology images and 1 billion unpaired text tokens, followed

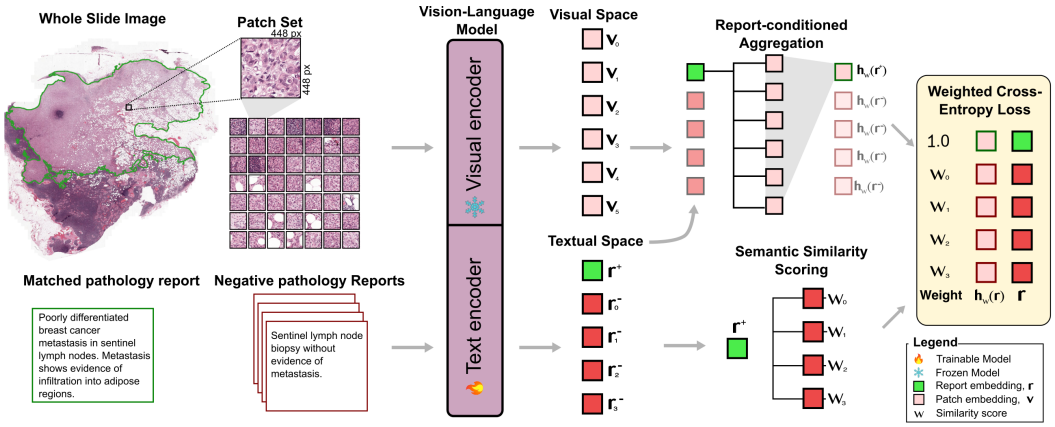

Figure 1: **Overview of LIMIT.** A frozen vision encoder extracts patch embeddings from the WSI. The text encoder embeds each candidate report as a cross-attention query over the patch bag, producing a report-conditioned slide representation. Only the text encoder is updated via a difficulty-weighted InfoNCE objective.

by contrastive alignment on Quilt-1M [4]. However, these patch-level VL encoders remain primarily trained on curated captions rather than the slide-level diagnostic reports routinely produced in clinical practice.

Several recent works instead leverage slide-level supervision from whole-slide images and pathology reports. PRISM, Prov-GigaPath, PathAlign, and TITAN learn slide-level representations by training dedicated aggregation modules or slide encoders on top of patch features, using report supervision, synthetic captions, or large-scale self-supervised objectives [15–17, 6]. These models establish the value of slide-report pairs for whole-slide representation learning, but their compressed slide-level embeddings can reduce native access to fine-grained patch-level visual-language alignment, which is important for patch preprocessing [18, 7] and post-hoc interpretability [8]. In contrast, LIMIT uses slide-level clinical reports to adapt a patch-level VL model, aiming to retain unified patch- and slide-level zero-shot capabilities without training a new slide encoder.

Cross-Attention Pooling in Vision-Language Models. Cross-attention is widely used to aggregate visual features conditioned on text or learned queries. CoCa uses cross-attention over ViT patch tokens [19], BLIP-2's Q-Former queries frozen visual features [20], and X-Pool uses text embeddings as queries over video frame features for text-video retrieval [21]. In computational pathology, MIL methods such as ABMIL, CLAM, and TransMIL learn attention-based aggregation over patch embeddings for slide-level prediction [22–24]. LIMIT differs by deriving the attention query from the pathology report text encoder and using this report-conditioned pooling directly as the alignment objective for locked-image text tuning.

Locked-image Tuning and Modality Alignment. Locked-image Tuning (LiT) freezes a pretrained vision encoder and updates only the text encoder, aligning language representations to an already robust visual space [25]. This is especially useful in pathology, where fine-tuning the visual backbone on noisy report supervision may risk degrading learned morphological features. LIMIT applies this principle to diagnostic text from 335,645 slide–text pairs, aligning the text encoder of a patch-level VL model to frozen pathology image features through weak slide-level supervision.

3 Methods

In this section, we describe Locked Image Multiple Instance PreTraining (LIMIT), our approach for weakly-supervised VL alignment in whole slide images (WSIs). LIMIT employs a locked-image tuning strategy, wherein the image encoder remains frozen while the text encoder is fine-tuned. To bridge the spatial disconnect between slide-level reports and patch-level visual features, LIMIT uses report embeddings as cross-attention queries to dynamically aggregate patch-level features into report-conditioned slide-level representations. The text encoder is optimized using a difficulty-weighted

InfoNCE objective that leverages inter-report semantic similarities to scale the influence of hard negatives.

3.1 Report-Guided Slide-Level Representation

A whole slide image W is divided into a set of M non-overlapping patches, $P = \{p_1, p_2, \ldots, p_M\}$. We extract patch embeddings using a frozen, pretrained vision encoder E_I, yielding a set of visual embeddings $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_M\}$, where $\mathbf{v}_i \in \mathbb{R}^d$.

To aggregate these dense patch representations into a single slide-level embedding, LIMIT introduces a text-driven attention mechanism conditioned on the pathology reports themselves. For a given pathology report R, we embed it using our trainable text encoder E_T to obtain the report vector $\mathbf{r} = E_T(R) \in \mathbb{R}^d$. We then compute the alignment between this report embedding and each patch embedding via a scaled dot-product similarity.

Because the aggregation is guided by the text, the resulting slide-level representation $\mathbf{h}_W(\mathbf{r}) \in \mathbb{R}^d$ becomes uniquely conditioned on the specific report being evaluated. It is computed as the softmax-weighted sum of the patch embeddings within the bag:

$$\alpha_i(\mathbf{r}) = \frac{\exp(\mathbf{v}_i^\top \mathbf{r}/\tau_{attn})}{\sum_{k=1}^{M} \exp(\mathbf{v}_k^\top \mathbf{r}/\tau_{attn})} \tag{1}$$

$$\mathbf{h}_W(\mathbf{r}) = \sum_{i=1}^{M} \alpha_i(\mathbf{r}) \mathbf{v}_i \tag{2}$$

where τ_{attn} is a temperature hyperparameter. This aggregation allows each candidate description to emphasize different regions of the same slide.

3.2 Difficulty-Weighted Contrastive Optimization

During training, each WSI W is paired with its corresponding ground-truth pathology report R^+. To construct a robust learning objective, we sample N additional slide-report pairs exclusively from the same organ type to serve as hard negatives. Together with the anchor pair, these form a bag of $K = N + 1$ paired examples, $\mathcal{B} = \{(W_i, R_i)\}_{i=1}^{K}$. Each report is embedded using the text encoder to obtain $\mathbf{r}_i = E_T(R_i)$. Every slide in the bag is matched against all K reports, with its own report serving as the positive and the remaining reports serving as negatives.

Because negative reports sampled from the same organ may exhibit high degrees of semantic overlap with the corresponding positive report (e.g., identical cancer subtype and grade), treating all negatives uniformly risks penalizing the model for pushing apart semantically valid alternative descriptions [26–28]. To address this, we compute the cosine similarity between each positive report and its negative reports to dynamically downweight the penalty applied to semantically overlapping negatives. Letting $\mathrm{sim}(\cdot, \cdot)$ denote this similarity, we define a linear-scale weighting factor w_{ij} for each mismatched pair:

$$w_{ij} = 1 - \frac{\mathrm{sim}(\mathbf{r}_i, \mathbf{r}_j) + 1}{2}, \qquad i \neq j. \tag{3}$$

The weights are computed from detached embeddings of the current trainable text encoder. Equivalently, $w_{ij} = \mathrm{stopgrad}[(1 - \mathrm{sim}(\mathbf{r}_i, \mathbf{r}_j))/2]$. Semantically similar negatives receive smaller weights, reducing penalties on potentially valid alternative descriptions. Gradients still flow through report-conditioned pooling and the slide–text alignment scores.

The model is optimized as a multi-class classification task, predicting the correct report out of the K candidates for every slide in the bag. Because our slide embeddings are report-dependent, we generate a unique, text-conditioned slide representation $\mathbf{h}_{W_i}(\mathbf{r}_j)$ for each slide-report combination. We compute the similarity between these conditioned representations and their respective text embeddings. The final training objective is a similarity-weighted InfoNCE loss averaged over all slides in the bag:

$$\mathcal{L} = -\frac{1}{K} \sum_{i=1}^{K} \log \frac{\exp(\mathbf{h}_{W_i}(\mathbf{r}_i)^\top \mathbf{r}_i/\tau_{nce})}{\exp(\mathbf{h}_{W_i}(\mathbf{r}_i)^\top \mathbf{r}_i/\tau_{nce}) + \sum_{\substack{j=1 \\ j \neq i}}^{K} w_{ij} \exp(\mathbf{h}_{W_i}(\mathbf{r}_j)^\top \mathbf{r}_j/\tau_{nce})}. \tag{4}$$

where $\tau_{nce} = 0.07$ is the contrastive temperature scalar. By freezing the image encoder and backpropagating $\mathcal{L}$ exclusively through E_T, LIMIT adapts the text representation to the fixed visual space, allowing for computation efficient large-scale VL alignment.

3.3 Inference and Evaluation

We distinguish patch-level classification from candidate-conditioned slide-level classification. For candidate labels $\mathcal{C} = \{c_1, \ldots, c_K\}$, let $\mathbf{c}_k = E_T(c_k)$ denote the adapted text embeddings.

Patch-level inference. Given a patch p with frozen visual embedding $\mathbf{v} = E_I(p)$, prediction uses direct cosine similarity:

$$\hat{y}_{\text{patch}} = \underset{k \in \{1,\ldots,K\}}{\arg\max}\ \text{sim}(\mathbf{v}, \mathbf{c}_k). \tag{5}$$

No slide-level pooling or downstream classifier is required.

Slide-level inference. For each candidate, we construct a distinct pooled representation $\mathbf{h}_W(\mathbf{c}_k)$ using the attention mechanism in Section 3.1. Prediction is

$$\hat{y}_{\text{slide}} = \underset{k \in \{1,\ldots,K\}}{\arg\max}\ \text{sim}(\mathbf{h}_W(\mathbf{c}_k), \mathbf{c}_k). \tag{6}$$

Candidate embeddings can be cached. For M patches, K candidates, and embedding dimension d, computing patch–candidate similarities and pooled representations requires $O(MKd)$ arithmetic; a materialized similarity matrix requires $O(MK)$ storage. This cost excludes visual feature extraction and text encoding. Empirical timing measurements are discussed in Appendix A.11.

3.4 Experiments

Datasets. We evaluate CONCH-Z on five datasets spanning slide- and patch-level zero-shot tasks; full dataset details are provided in **Section A.2**. For slide-level evaluation, we use **TCGA-OT** [6], a cancer subtyping benchmark with 46 OncoTree disease labels across multiple anatomical sites, which we reduce to 45 classes after post-processing (See **Appendix A.2**) and **SlideChat-HistAI** [9], a diagnostic subtyping task across 3 organs and 1,396 slides. For both slide-level tasks, we formulate this as an organ-specific multiple-choice task, mirroring clinical workflows where tissue origin is known *a priori*. For each WSI, valid diagnostic options are restricted to its corresponding organ. For patch-level evaluation, we use **SPIDER** [29], a four-organ expert-annotated benchmark with 408,235 annotated patches covering skin, breast, colorectal, and thorax tissue, evaluated in both fine-grained form (SPIDER-F; 13–24 classes per organ) and a coarse unified taxonomy (SPIDER-C). We also evaluate tumor detection on **CAMELYON-16** using CAMELYON+ annotations [30, 31], yielding 206,800 labeled patches from 48 slides, and out-of-domain cancer detection (2 classes) and tissue classification (11 classes) on **CATCH** [32], comprising 1,251,281 patches from canine cutaneous biopsies.

Zero-Shot and Vision-Language Baselines. We compare the supervised models against state-of-the-art zero-shot computational pathology models at the patch-level (MI-Zero [5], CONCHv1.5[3], MUSK [4], QuiltNet [14], and PLIP [13]), and at the slide-level (TITAN [6], PRISM [15]). Patch-level VL models use the inference pipeline described in **Section 3.3**. Slide-level foundation models use their respective slide representations for candidate-text scoring.

Supervised Baselines. To evaluate our zero-shot approach against fully supervised baselines, we utilize ABMIL [22], CLAM [23], DSMIL [33], TransMIL [24], and WiKG [34]. Models are trained on the TCGA-OT training split using a standard 45-way cross-entropy objective. During inference, to ensure a fair comparison with our zero-shot organ-specific forumalation, we constrain the supervised predictions by applying an organ-specific mask. The final prediction is designated as the subtype with the highest predicted probability strictly within the slide's known organ category. We follow the standard training recipe for MIL models in pathology [35, 11], further described in **Section A.1**.

Pretraining and Implementation Details. We train LIMIT on Mass-340K [6], comprising 335,645 WSIs from Mass General Brigham and GTEx. We use 448×448 pixel patches unless otherwise specified. LIMIT is conducted for 3 epochs with training batches of 5 slide–report pairs from the same organ and a learning rate of 1×10^{-5} with weight decay 1×10^{-5} using pre-computed patch-level

features from the Trident library[36]. We use the AdamW optimizer with a cosine annealing learning rate schedule. Within each batch, all 5 slides are scored against all 5 reports, yielding 25 slide–report combinations. Each slide has one matching report and 4 negative reports; the loss is averaged over the 5 slides. A 1% validation split is held out for selecting the best epoch. Training text consists of normalized diagnostic phrases extracted from reports with a locally-hosted Qwen3-14B model [37] to remove residual identifying information and extract diagnostic content, excluding details such as laterality, staining, and ancillary marker results (detailed in **Section A.2**). For zero-shot target labels, we follow the CPath convention of prepending the tissue type to the class label [38, 5, 39].

4 Results

4.1 Zero-Shot Patch-Level Classification

We begin our evaluation by investigating the capacity of CONCH-Z to perform detailed, multi-class patch-level classification. To identify whether our slide-level training regime successfully aligns slide-level diagnostic text with fine-grained microscopic morphologies, we evaluate the model on large-scale patch-level benchmarks across both human (SPIDER) and canine (CATCH) tissues. Despite being optimized exclusively on slide-level reports, CONCH-Z demonstrates highly competitive zero-shot performance across varied tissue types and species.

Table 1: **Patch-level zero-shot.** Performance on patch-level tasks for C16 (AUC), SPIDER (balanced accuracy), and CATCH (balanced accuracy). Number of classes shown in parentheses. Best performance in **bold**; second-best underlined. Standard deviation according to 1,000 bootstrap trials.

Dataset	Organ	PLIP[13]	QuiltNet[14]	Musk[4]	Mi-Zero[5]	CONCH[3]	CONCH-Z
C16 (2)	Lymph Node (2)	88.4 ± 0.1	91.7 ± 0.1	82.7 ± 0.2	59.9 ± 0.2	90.3 ± 0.2	**96.7 ± 0.1**
SPIDER-C	Skin (9)	38.0 ± 0.3	30.8 ± 0.3	**62.1 ± 0.3**	39.2 ± 0.3	50.8 ± 0.3	60.8 ± 0.3
	Colorectal (8)	43.5 ± 0.3	37.7 ± 0.4	43.5 ± 0.3	47.7 ± 0.4	44.0 ± 0.4	**54.8 ± 0.4**
	Breast (7)	32.8 ± 0.3	28.2 ± 0.5	57.8 ± 0.4	50.6 ± 0.4	57.9 ± 0.4	**70.2 ± 0.3**
	Thorax (7)	25.0 ± 0.4	23.6 ± 0.4	41.7 ± 0.4	39.2 ± 0.5	46.2 ± 0.4	**49.4 ± 0.4**
	Overall	34.8 ± 0.2	30.1 ± 0.2	51.3 ± 0.2	44.2 ± 0.2	49.7 ± 0.2	**58.8 ± 0.2**
SPIDER-F	Skin (24)	32.1 ± 0.2	17.4 ± 0.2	50.8 ± 0.2	33.0 ± 0.2	42.3 ± 0.2	**53.3 ± 0.2**
	Colorectal (13)	30.9 ± 0.3	28.4 ± 0.3	34.5 ± 0.3	36.7 ± 0.3	38.5 ± 0.3	**48.6 ± 0.4**
	Breast (18)	17.3 ± 0.2	11.3 ± 0.3	41.7 ± 0.4	36.5 ± 0.4	46.9 ± 0.4	**52.7 ± 0.4**
	Thorax (14)	22.0 ± 0.2	21.5 ± 0.2	49.2 ± 0.3	33.4 ± 0.3	43.8 ± 0.3	**50.7 ± 0.3**
	Overall	25.6 ± 0.1	19.7 ± 0.1	44.1 ± 0.2	34.9 ± 0.1	42.9 ± 0.1	**51.3 ± 0.2**
CATCH	Skin-F (11)	19.2 ± 0.0	25.3 ± 0.0	44.2 ± 0.1	41.4 ± 0.1	23.4 ± 0.0	**45.1 ± 0.0**
	Skin-C (2)	79.4 ± 0.0	71.3 ± 0.0	74.0 ± 0.0	61.2 ± 0.0	72.4 ± 0.0	**83.7 ± 0.0**
	Overall	49.3 ± 0.0	48.3 ± 0.0	59.1 ± 0.0	51.3 ± 0.0	47.9 ± 0.0	**64.4 ± 0.0**

On the patch-level benchmarks, CONCH-Z demonstrates broad improvements over existing VL models, establishing new state-of-the-art performance across nearly all evaluated sub-tasks. On the coarse-grained SPIDER-C dataset, CONCH-Z establishes achieves top accuracies in Colorectal (54.8%), Breast (70.2%), and Thorax (49.4%), while remaining competitive on Skin (60.8%, second to Musk), substantially outperforming models pretrained on patch-caption pairs. A similar trend is observed in the fine-grained SPIDER-F task, where CONCH-Z leads across all organ types, achieving top marks in Skin (53.3%), Colorectal (48.6%), Breast (52.7%), and Thorax (50.7%) to deliver an overall balanced accuracy of 51.3%. Similarly, we observe consistent improvements on the canine CATCH benchmark, where CONCH-Z achieves the highest performance in both coarse-grained cancer detection (83.7%) and fine-grained subtyping (45.1%). Our overall results confirm that locked-image tuning on slide-level reports successfully enhances microscopic visual-semantic alignment.

4.2 Zero-Shot Tumor Detection and Spatial Interpretability

To assess grounding of slide-level text to patch-level visual features without explicit spatial supervision or complex MIL aggregators, we evaluate on the CAMELYON-16 (C16) benchmark which contains pixel-level annotations for breast cancer metastasis detection.

CONCH-Z achieves an overall zero-shot AUC of 96.7%, outperforming both QuiltNet (91.7%) and CONCHv1.5 (90.3%). Beyond slide-level classification, this fine-grained semantic alignment enables

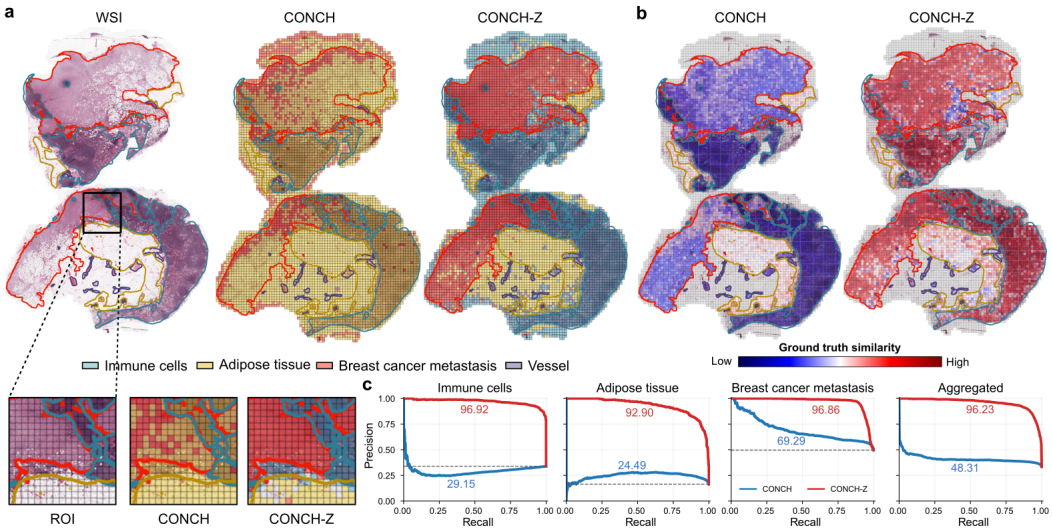

Figure 2: **Zero-shot breast cancer metastasis detection on the CAMELYON16 dataset**. **a)** 4-way patch classification performed according to maximum cosine similarity. Metastasis annotations were provided by Camelyon+, while the remaining annotations were generated by an in-house pathologist. **b)** Model score for the ground truth class. Red indicates the ground truth class received high predicted scores, while blue indicates low scores for the predicted class. Scores are normalized per-patch, and heatmap range is determined per-model. **c)** Area Under the Precision-Recall Curve for the key tissue categories for CONCHv1.5 (blue) and CONCH-Z (red).

inherent spatial interpretability. As illustrated in **Figure 2**, when prompted with key diagnostic terms, the aligned text encoder achieves substantial overlap with pathologist-annotated regions of metastasis, adipose tissue, and lymphocytes. Visualizations for other models are in **Figure A1**.

4.3 Zero-Shot Slide-Level Classification

We next evaluate the representational capacity of our optimized text encoder on slide-level subtyping tasks using the TCGA-OT (n=11,186) and SlideChat-HistAI (n=1,396) datasets. CONCH-Z achieves the highest zero-shot performance across both benchmarks, outperforming heavily parameterized slide-level foundation models as well as established patch-level baselines (**Table 2**).

On the TCGA-OT benchmark, CONCH-Z achieves an overall zero-shot balanced accuracy of 78.8%, representing an absolute improvement of 21.4 percentage points over its baseline VL backbone (CONCH, 57.4%) and an improvement of 5.7 percentage points over the leading slide-level foundation model, TITAN (73.1%, compared to PRISM at 59.4%). We visualize this stronger alignment of the text and visual features in **Fig. A2**.

This performance gap persists on the HistAI dataset, which poses a 3-class classification task for breast, colorectum, and skin cancers. CONCH-Z achieves an overall balanced accuracy of 54.7%, outperforming the next best patch model (Musk, 44.6%) and slide model (PRISM, 50.2%) by a wide margin. These results support improved alignment for organ-conditioned diagnostic classification after adapting the text encoder with slide-level diagnostic supervision.

4.4 Comparison with Supervised Baselines and Learned Aggregation

We next sought to contextualize the performance of zero-shot classification and popular supervised approaches. We evaluate CONCH-Z against a suite of fully supervised MIL architectures on the TCGA-OT test set (1,348 slides). The supervised models (ABMIL, TransMIL, DSMIL, WiKG, and CLAM) were trained for 45-way classification on 8,226 samples from the official TCGA-OT splits, with melanoma and cutaneous melanoma grouped into a single class due to label noise (See **Appendix A.2**). To ensure a fair comparison with our zero-shot formulation, we select the highest logit within-organ prediction during the evaluation of these supervised models.

Table 2: **Slide-level zero-shot with organ-restricted candidates.** HistAI and TCGA-OT per-organ balanced accuracy. TCGA results here use the full evaluation cohort (11,186 slides), whereas Table 3 uses only the 1,348-slide test split. Best **bold**; second-best underlined. Standard deviation reported according to 1,000 bootstrap trials.

Dataset	Organ	Patch foundation model						Slide foundation model	
		PLIP[13]	QuiltNet[14]	Musk[4]	Mi-Zero[5]	CONCH[3]	CONCH-Z	TITAN[6]	PRISM[15]
HistAI	Breast	41.6 ± 3.6	32.3 ± 1.3	49.5 ± 3.6	44.7 ± 2.8	34.0 ± 1.4	47.6 ± 3.6	46.1 ± 3.5	**49.7 ± 2.7**
	Colorectum	28.4 ± 2.2	33.8 ± 3.6	36.2 ± 3.6	31.4 ± 2.9	32.9 ± 3.1	**51.4 ± 4.2**	33.2 ± 2.9	44.3 ± 3.8
	Skin	47.6 ± 3.1	33.3 ± 0.0	48.1 ± 3.3	46.0 ± 3.1	53.1 ± 3.7	**65.0 ± 3.1**	58.9 ± 3.0	56.7 ± 3.7
	Overall	39.2 ± 1.7	33.1 ± 1.3	44.6 ± 2.1	40.7 ± 1.7	40.0 ± 1.8	**54.7 ± 2.2**	46.1 ± 1.7	50.2 ± 2.0
TCGA	Breast	48.8 ± 0.6	7.2 ± 0.9	62.8 ± 1.6	49.6 ± 1.8	50.0 ± 0.6	**89.2 ± 1.2**	77.7 ± 1.7	81.9 ± 1.3
	Lung	52.5 ± 3.1	27.8 ± 4.4	68.1 ± 3.2	31.9 ± 4.2	47.4 ± 3.9	74.2 ± 4.0	**95.2 ± 0.9**	57.1 ± 3.8
	Brain	17.9 ± 1.2	33.4 ± 6.5	15.6 ± 1.5	14.7 ± 1.6	39.2 ± 5.0	**44.4 ± 4.6**	33.0 ± 5.5	26.9 ± 6.0
	Kidney	66.9 ± 1.8	58.4 ± 1.9	57.2 ± 1.5	75.5 ± 1.7	77.9 ± 2.6	**96.4 ± 0.7**	83.0 ± 2.0	67.0 ± 3.8
	Thyroid	25.4 ± 1.9	45.2 ± 2.6	52.3 ± 2.6	67.7 ± 2.2	76.2 ± 2.4	**82.9 ± 2.1**	74.6 ± 1.8	62.9 ± 2.6
	Bowel	39.2 ± 5.4	40.3 ± 3.1	57.5 ± 4.3	39.6 ± 7.4	58.2 ± 4.7	**74.7 ± 2.1**	64.7 ± 2.5	51.6 ± 4.8
	Uterus	34.0 ± 2.7	50.2 ± 0.2	45.1 ± 2.6	35.4 ± 2.3	46.8 ± 1.2	**82.1 ± 2.1**	80.0 ± 1.5	65.4 ± 2.6
	Other	38.2 ± 3.3	35.4 ± 3.4	49.5 ± 3.5	37.6 ± 3.4	63.3 ± 3.4	**86.7 ± 2.8**	76.3 ± 2.6	62.4 ± 3.2
	Overall	40.4 ± 1.6	37.2 ± 1.4	51.0 ± 1.4	44.0 ± 1.5	57.4 ± 1.7	**78.8 ± 1.7**	73.1 ± 0.9	59.4 ± 1.7

Table 3: **TCGA-OT supervised baselines** Supervised baselines and linear probes were trained on 8,226 samples and evaluated on 1,348 samples. Balanced accuracy is reported for all cases (± standard deviation). **Overall** represents the average balanced accuracy across all organs.

Type	Model	Breast	Lung	Brain	Kidney	Thyroid	Bowel	Uterus	Other	Overall
Supervised	ABMIL[22]	**91.0 ± 3.6**	88.5 ± 2.6	26.5 ± 4.5	89.4 ± 3.2	57.3 ± 3.0	**63.1 ± 10.1**	81.8 ± 5.1	89.8 ± 1.0	73.4 ± 0.9
	TransMIL[24]	63.6 ± 5.5	**90.8 ± 2.4**	31.8 ± 4.8	87.7 ± 3.5	68.1 ± 3.4	58.6 ± 10.6	78.6 ± 5.5	88.8 ± 1.2	71.0 ± 1.1
	DSMIL[33]	80.9 ± 5.3	87.9 ± 3.0	32.8 ± 5.5	88.7 ± 3.5	48.8 ± 1.3	59.6 ± 10.7	**84.6 ± 4.3**	85.5 ± 2.7	71.1 ± 2.0
	CLAM[23]	88.9 ± 4.0	89.7 ± 2.6	**36.8 ± 5.8**	82.0 ± 4.0	50.2 ± 1.5	52.2 ± 10.4	84.1 ± 4.2	88.2 ± 1.3	71.5 ± 1.2
	WiKG[34]	88.5 ± 4.6	88.0 ± 2.9	36.6 ± 5.9	88.2 ± 3.7	47.5 ± 0.4	56.9 ± 10.4	73.1 ± 5.7	88.0 ± 1.2	70.9 ± 1.1
	$\text{CONCH}_{\text{lin}}$[3]	83.9 ± 5.2	87.1 ± 3.1	30.1 ± 7.5	89.7 ± 3.3	**80.4 ± 3.2**	62.9 ± 9.2	79.2 ± 5.5	88.9 ± 1.1	75.3 ± 1.1
	$\text{CONCH-Z}_{\text{lin}}$	83.9 ± 5.2	86.8 ± 3.2	33.4 ± 7.9	**90.7 ± 3.1**	**80.4 ± 3.2**	62.9 ± 9.2	78.9 ± 5.5	**90.8 ± 1.1**	**76.0 ± 1.1**
Zero-shot	TITAN[6]	83.5 ± 5.1	91.4 ± 2.5	32.2 ± 6.4	77.1 ± 4.2	83.7 ± 2.5	66.0 ± 9.3	83.2 ± 3.5	61.7 ± 1.8	72.4 ± 1.4
	PRISM[15]	82.0 ± 4.9	60.6 ± 9.7	20.4 ± 5.9	58.4 ± 6.3	66.5 ± 4.0	65.6 ± 9.1	62.6 ± 6.0	74.2 ± 1.9	61.3 ± 1.6
	MI-Zero[5]	52.6 ± 5.0	45.8 ± 13.2	19.2 ± 5.9	73.4 ± 5.2	73.4 ± 3.2	47.4 ± 10.9	48.9 ± 4.0	40.9 ± 3.8	50.2 ± 2.6
	PLIP[13]	50.0 ± 0.0	75.0 ± 6.3	12.5 ± 4.7	65.4 ± 5.2	27.7 ± 4.2	51.0 ± 10.7	41.9 ± 8.6	42.9 ± 4.0	45.8 ± 2.7
	QuiltNet[14]	19.6 ± 5.9	46.0 ± 13.1	20.0 ± 6.9	57.5 ± 5.7	52.0 ± 3.8	58.5 ± 10.2	51.4 ± 5.7	26.0 ± 2.7	41.4 ± 2.1
	Musk[4]	73.5 ± 5.0	69.0 ± 7.7	12.5 ± 4.5	58.0 ± 4.8	62.6 ± 3.9	**67.4 ± 8.7**	52.4 ± 6.2	57.3 ± 3.8	56.6 ± 2.7
	CONCH[3]	49.8 ± 2.8	80.3 ± 4.9	**62.4 ± 5.1**	79.0 ± 4.5	85.7 ± 2.9	54.7 ± 8.0	50.0 ± 0.0	70.0 ± 3.5	66.5 ± 2.4
	CONCH-Z	**90.7 ± 4.2**	**91.5 ± 2.6**	33.7 ± 6.8	**89.4 ± 3.6**	**92.4 ± 2.1**	66.1 ± 9.6	**83.2 ± 4.8**	**92.8 ± 1.1**	**80.0 ± 1.0**

Furthermore, we study the efficacy of our zero-shot models against learned aggregation schemes with an additional linear probe baseline. In this setup, we utilize the base patch encoders (CONCH and CONCH-Z) paired with the text query "cancer" to serve as a zero-shot spatial aggregator. Subsequently, a linear probe ($\text{CONCH}_{\text{lin}}$and $\text{CONCH-Z}_{\text{lin}}$) is trained on these aggregated representations to produce the final prediction. Remarkably, this fixed, zero-shot aggregation performs competitively with supervised aggregators, with $\text{CONCH-Z}_{\text{lin}}$achieving 76.0% overall balanced accuracy, outperforming the best-performing traditional MIL model (ABMIL, 73.4%) by 2.6 percentage points within the evaluated supervised comparisons. Meanwhile, using CONCH-Z for end-to-end zero-shot classification achieves an overall balanced accuracy of 80.0% (**Table 3**), representing a gain of 6.6 percentage points over the best supervised MIL approach (ABMIL, 73.4%) and 4.0 percentage points over the fixed-query linear probe, $\text{CONCH-Z}_{\text{lin}}$(76.0%). These findings highlight the importance of query-aware aggregation paired with zero-shot classification.

4.5 Generalization across Models and Training Corpora

To evaluate whether LIMIT functions as a general-purpose adaptation framework beyond CONCHv1.5, we applied our training procedure to PLIP, QuiltNet, and MUSK using a randomly selected subset of 25,000 slide–report pairs from our internal corpus. Although these baseline models differ substantially in their underlying architectures and initial pretraining objectives, adapting them with LIMIT yielded consistent downstream improvements across the benchmark suite, enhancing performance on all 17 evaluated tasks for both PLIP and MUSK, and on 15 of the 17 tasks for QuiltNet (Appendix A.7). These consistent gains indicate that LIMIT effectively transfers diagnostic

signal from slide-level reports into patch-level representations across diverse model families and pretraining paradigms.

In addition to evaluating architectural flexibility, we also study whether LIMIT requires proprietary clinical data to succeed or if it can be effectively trained using smaller, publicly accessible repositories. To test this, we adapted CONCHv1.5 using 10,108 publicly available whole-slide images and their corresponding pathology reports from TCGA. When evaluated on the external SPIDER and CATCH benchmarks, this public-data adaptation improved balanced accuracy across all ten constituent tasks, achieving an equal-weight average gain of 6.6 percentage points over the unadapted baseline model. Notably, this performance closely approaches the 7.0 percentage point average improvement achieved when adapting the same backbone on the larger set of 25,000 internal slide–report pairs across the identical evaluation tasks (Appendix A.8). We conclude that LIMIT does not depend on massive, institution-specific datasets, with publicly available slide–report pairs providing sufficient supervisory signal to drive substantial improvements in cross-dataset transfer.

4.6 Ablations

We evaluate the influence of various training-time and inference-time decisions to assess the robustness of our approach in **Table 4**, which reports the balanced accuracy averaged across TCGA-OT and HistAI tasks after altering various train-time and test-time approaches.

Pretrain Size: We consider three sizes of the pretraining set: the full 335,645 samples, 150k samples, and 50k samples, in order to provide greater insight into how each design decision influences performance at various pretrain scales. Following LIMIT pretraining, we observe steady performance scaling with data volume across evaluation configurations, with balanced accuracy increasing from 62.7% at 50k to 66.8% with the full corpus for our primary model.

Negative selection: We investigate various approaches for selecting N negative reports during training. While our standard training approach only retrieves reports from the same organ, allowing random organ selection causes a 2.9% relative drop with the full corpus and a 3.5% drop at 50k (1.9 and 2.2 percentage points, respectively). We also vary the number of negative reports retrieved from 1 to 8. While CONCH-Z samples 4 negative reports per batch, performance remains robust when sampling larger numbers of negatives ($N = 6$ or 8) with minimal degradation (up to a 1.8% relative drop with the full corpus), but degrades more noticeably when using fewer negatives ($N = 2$ or 1, falling by up to 7.1% relatively with the full corpus), indicating that a sufficient pool of hard in-organ negatives is essential for effective contrastive alignment.

Loss: We remove the weighted loss scaling, such that all reports are equally weighted regardless of embedding similarity to the positive pair report. Removing this weighted loss causes a substantial drop across all regimes, decreasing performance by 8.1 percentage points (a 12.2% relative drop) with the full corpus, which confirms that weight scaling by report similarity is critical to prevent penalizing semantically congruent reports. Replacing the InfoNCE loss with weighted SigLIT loss [40], a prominent contrastive loss for LiT, yields performance closely tracking InfoNCE across all pretraining regimes (e.g., 66.3% versus 66.8% with the full corpus), suggesting that LIMIT pretraining is robust to the choice of contrastive objective.

Finetune: We finetune with low-rank adaptation (LoRA) [41] rather than full weight finetuning. LoRA performance gains remain lower across all dataset scales (e.g., 64.1% versus 66.8% with the full corpus), indicating that updating the full set of text representation weights is advantageous for fine-grained clinical domain transfer.

Pooling method: Rather than pool CONCH-Z with weighted similarity to each text option at inference-time, we perform mean pooling and random pooling (reported as the average of ten trials per task). Replacing LIMIT pooling with mean or random pooling leads to noticeable relative drops of 3–8% and 11%, highlighting the importance of content-aware aggregation. Aggregating the slide from top-10 and top-100 patches yields comparable performance to weighted aggregation of all patches with LIMIT, but results vary across training scales. See **Appendix A.4** for further details on ablation implementation.

Synonym evaluation: We evaluate sensitivity to terminology and phrasing using five cancer-related terms and three sentence templates (Table A4). CONCH-Z achieves higher average performance

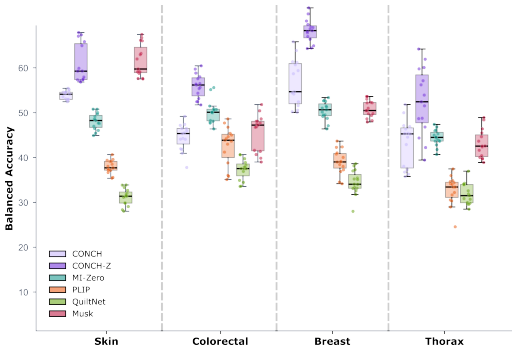

Figure 3: SPIDER-C tissue subtyping performance when the cancer class prompt is substituted with every combination of five cancer-related synonyms and three sentence templates. Each point represents performance with a single synonym-template pair.

Table 4: Ablation studies. Effect of various design choices for training and inference on balanced accuracy, averaged across TCGA-OT (n=11,186) and HistAI (n=1,396)

			Pretrain Size		
	Ablation	**Setting**	**Full**	**150k**	**50k**
Training	**Full model**	Ours	66.8%	64.8%	62.7%
	Negative selection	Rand. organ	64.7%	64.0%	60.5%
		$N = 8$	65.7%	62.0%	62.1%
		$N = 6$	65.4%	64.8%	63.0%
		$N = 2$	63.9%	62.6%	61.7%
		$N = 1$	61.9%	62.0%	60.5%
	Loss	No weight	58.5%	59.1%	59.1%
		SigLIT [40]	66.3%	65.1%	61.5%
	Finetune	LoRA[41]	64.1%	63.7%	61.8%
Inference	Pooling method	Mean pool	61.5%	59.5%	60.6%
		Rand. pool	59.0%	57.5%	55.7%
		Top-10	65.2%	64.6%	60.4%
		Top-100	66.7%	63.0%	62.7%

across the evaluated prompt variants, while remaining sensitive to wording, particularly for skin and thorax.

Visual adaptation and initialization. On a 25,000-pair subset, we compared text-only LIMIT with LIMIT-V, which uses frozen precomputed features to select the top-k patches for each diagnostic query and re-encodes selected patches through a trainable vision encoder. With $k = 64$, LIMIT-V required 55.3 GB peak GPU memory, compared with 4.3 GB for the corresponding locked-image measurement, and required 44 hours for one training epoch on an A100 GPU. Its SPIDER performance was mixed relative to text-only adaptation (Appendix A.9). This supports the practical performance–compute tradeoff of frozen visual features under the evaluated setup.

Finally, we investigated whether LIMIT requires a pretrained text encoder for fine-tuning or if it can establish cross-modal alignment entirely from scratch on the 25,000-pair internal subset. When we paired the frozen CONCHv1.5 visual backbone with a randomly initialized text encoder, training yielded modest average gains of $+5.90$ and $+6.33$ percentage points over the untrained random baseline on SPIDER-F and SPIDER-C, respectively, falling short of the $+32.70$ and $+33.90$ percentage point improvements achieved when adapting the pretrained text encoder (CONCH-Z @25k; Table A11). This gap indicates that slide-level reports alone provide insufficient localized supervision to simultaneously learn both semantic text representations and spatial patch grounding from random weights. However, when we guided training using pooled queries derived from the frozen pretrained CONCHv1.5 text encoder, average downstream gains nearly tripled to $+16.70$ percentage points on SPIDER-F and $+20.93$ percentage points on SPIDER-C (Appendix A.10). We conclude that existing VL prior is essential for directing attention toward diagnostically relevant tissue regions during slide-level adaptation. Together, these findings demonstrate that LIMIT is best suited as a post-hoc fine-tuning framework rather than a method for pretraining from scratch.

Limitations: LIMIT adapts only the text encoder and cannot create visual features missing from the frozen backbone. Report-conditioned weighted pooling also lacks an explicit model of spatial relationships across regions, which may limit tasks requiring global architectural assessment. The method assumes that normalized diagnostic text describes evidence present in at least some slide patches. Reports describing other specimens, ancillary testing, or findings absent from the scanned section can violate this assumption. Preprocessing may also remove useful context or introduce errors.

Conclusion: We introduce LIMIT, a locked-image tuning framework that aligns patch-level pathology VL models with slide-level clinical reports through report-conditioned attention. Using LIMIT, CONCH-Z improves zero-shot patch- and slide-level classification across diverse pathology benchmarks while avoiding the parameter and compute overhead of training a dedicated slide encoder. Overall, these results suggest that routine diagnostic reports provide a scalable and underused source of supervision for strengthening visual-semantic alignment in computational pathology.

References

[1] Andrew H. Song, Guillaume Jaume, Drew F. K. Williamson, Ming Y. Lu, Anurag Vaidya, Tiffany R. Miller, and Faisal Mahmood. Artificial Intelligence for Digital and Computational Pathology, December 2023. URL http://arxiv.org/abs/2401.06148.

[2] Jana Lipkova, Tiffany Y. Chen, Ming Y. Lu, Richard J. Chen, Maha Shady, Mane Williams, Jingwen Wang, Zahra Noor, Richard N. Mitchell, Mehmet Turan, Gulfize Coskun, Funda Yilmaz, Derya Demir, Deniz Nart, Kayhan Basak, Nesrin Turhan, Selvinaz Ozkara, Yara Banz, Katja E. Odening, and Faisal Mahmood. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. *Nature Medicine*, 28(3):575–582, March 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01709-2. URL https://www.nature.com/articles/s41591-022-01709-2.

[3] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, March 2024. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-024-02856-4. URL https://www.nature.com/articles/s41591-024-02856-4.

[4] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Kun-Hsing Yu, Sierra Willens, Francesca Maria Olguin, Jeffrey J. Nirschl, Joel Neal, Maximilian Diehn, Sen Yang, and Ruijiang Li. A vision–language foundation model for precision oncology. *Nature*, 638(8051): 769–778, February 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08378-w. URL https://www.nature.com/articles/s41586-024-08378-w.

[5] Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual Language Pretrained Multiple Instance Zero-Shot Transfer for Histopathology Images, June 2023. URL http://arxiv.org/abs/2306.07831. arXiv:2306.07831 [cs.CV].

[6] Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, Drew F. K. Williamson, Harry Robertson, Bowen Chen, Cristina Almagro-Pérez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Christina S. Chen, Daisuke Komura, Akihiro Kawabe, Mieko Ochi, Shinya Sato, Tomoyuki Yokose, Yohei Miyagi, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31(11):3749–3761, November 2025. ISSN 1546-170X. doi: 10.1038/s41591-025-03982-3. URL https://www.nature.com/articles/s41591-025-03982-3.

[7] Wentao Huang, Weimin Lyu, Peiliang Lou, Qingqiao Hu, Xiaoling Hu, Shahira Abousamra, Wenchao Han, Ruifeng Guo, Jiawei Zhou, Chao Chen, and Chen Wang. Act Like a Pathologist: Tissue-Aware Whole Slide Image Reasoning, March 2026. URL http://arxiv.org/abs/2603.00667. arXiv:2603.00667 [cs].

[8] Daniel Shao, Joel Runevic, Richard J. Chen, Drew FK Williamson, Ahrong Kim, Andrew H. Song, and Faisal Mahmood. Mixture of Mini Experts: Overcoming the Linear Layer Bottleneck in Multiple Instance Learning. October 2025. URL https://openreview.net/forum?id=S5Io33pc78.

[9] Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. SlideChat: A Large Vision-Language Assistant for Whole-Slide Pathology Image Understanding, March 2025. URL http://arxiv.org/abs/2410.11761. arXiv:2410.11761 [cs] version: 3.

[10] Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, and Linlin Shen. WSI-LLaVA: A Multimodal Large Language Model for Whole Slide Image, August 2025. URL http://arxiv.org/abs/2412.02141. arXiv:2412.02141 [cs].

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

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

[13] 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, September 2023. ISSN 1546-170X. doi: 10.1038/s41591-023-02504-3. URL https://www.nature.com/articles/s41591-023-02504-3.

[14] Wisdom O Ikezogwo, Mehmet S Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir S Mohammed, Pavan K Anand, Ranjay Krishna, and Linda G Shapiro. Quilt-1M: One Million Image-Text Pairs for Histopathology.

[15] George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, Matthew Hanna, Julian Viret, Neil Tenenholtz, James Hall, Nicolo Fusi, Razik Yousfi, Peter Hamilton, A Moye, Eugene Vorontsov, Siqi Liu, and Thomas J Fuchs. PRISM: A MULTI-MODAL GENERATIVE FOUNDATION MODEL FOR SLIDE-LEVEL HISTOPATHOLOGY.

[16] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, June 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-07441-w. URL https://www.nature.com/articles/s41586-024-07441-w.

[17] Faruk Ahmed, Andrew Sellergren, Lin Yang, Shawn Xu, Boris Babenko, Abbi Ward, Niels Olson, Arash Mohtashamian, Yossi Matias, Greg S. Corrado, Quang Duong, Dale R. Webster, Shravya Shetty, Daniel Golden, Yun Liu, David F. Steiner, and Ellery Wulczyn. PathAlign: A vision-language model for whole slide images in histopathology, June 2024. URL http://arxiv.org/abs/2406.19578. arXiv:2406.19578 [cs].

[18] Junxian Wu, Minheng Chen, Xinyi Ke, Tianwang Xun, Xiaoming Jiang, Hongyu Zhou, Lizhi Shao, and Youyong Kong. Learning Heterogeneous Tissues with Mixture of Experts for Gigapixel Whole Slide Images. pages 5144–5153, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_Learning_Heterogeneous_Tissues_with_Mixture_of_Experts_for_Gigapixel_Whole_CVPR_2025_paper.html.

[19] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=Ee277P3AYC.

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[21] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5006–5015, 2022.

[22] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning, June 2018. URL http://arxiv.org/abs/1802.04712.

[23] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, June 2021. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w. URL https://www.nature.com/articles/s41551-020-00682-w.

[24] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification, October 2021. URL http://arxiv.org/abs/2106.00908.

[25] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-Shot Transfer with Locked-image text Tuning, June 2022. URL http://arxiv.org/abs/2111.07991. arXiv:2111.07991 [cs].

[26] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing Vision-Language Models via Biased Prompts, May 2023. URL http://arxiv.org/abs/2302.00070. arXiv:2302.00070 [cs.LG].

[27] Tri Huynh, Simon Kornblith, Matthew R. Walter, Michael Maire, and Maryam Khademi. Boosting Contrastive Self-Supervised Learning with False Negative Cancellation, January 2022. URL http://arxiv.org/abs/2011.11765. arXiv:2011.11765 [cs.CV].

[28] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.256. URL https://aclanthology.org/2022.emnlp-main.256/.

[29] Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. SPIDER: A COMPREHENSIVE MULTI-ORGAN SUPERVISED PATHOLOGY DATASET AND BASELINE MODELS.

[30] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, December 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.14585. URL https://doi.org/10.1001/jama.2017.14585.

[31] Xitong Ling, Yuanyuan Lei, Jiawen Li, Junru Cheng, Wenting Huang, Tian Guan, Jian Guan, and Yonghong He. Comprehensive Benchmark Dataset for Pathological Lymph Node Metastasis in Breast Cancer Sections. *Scientific Data*, 12(1):1381, August 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-05586-5.

[32] Frauke Wilm, Marco Fragoso, Christian Marzahl, Jingna Qiu, Chloé Puget, Laura Diehl, Christof A. Bertram, Robert Klopfleisch, Andreas Maier, Katharina Breininger, and Marc Aubreville. Pan-tumor CAnine cuTaneous Cancer Histology (CATCH) dataset. *Scientific Data*, 9(1):588, September 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01692-w. URL https://www.nature.com/articles/s41597-022-01692-w.

[33] Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning, April 2021. URL http://arxiv.org/abs/2011.08939.

[34] Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11323–11332, 2024.

[35] Daniel Shao, Richard J. Chen, Andrew H. Song, Joel Runevic, Ming Y. Lu, Tong Ding, and Faisal Mahmood. Do Multiple Instance Learning Models Transfer? June 2025. URL https://openreview.net/forum?id=hfLqdquVt3.

[36] Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating Data Processing and Benchmarking of AI Models for Pathology, February 2025. URL http://arxiv.org/abs/2502.06750. arXiv:2502.06750 [cs].

[37] Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[38] Cagla Deniz Bahadir, Gozde B Akar, and Mert R Sabuncu. LLM-Generated Rewrite and Context Modulation for Enhanced Vision Language Models in Digital Pathology.

[39] Vasudev Sharma, Ahmed Alagha, Abdelhakim Khellaf, Vincent Quoc-Huy Trinh, and Mahdi S. Hosseini. Investigating Zero-Shot Diagnostic Pathology in Vision-Language Models with Efficient Prompt Design, April 2025. URL http://arxiv.org/abs/2505.00134. arXiv:2505.00134 [cs.CV].

[40] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training, September 2023. URL http://arxiv.org/abs/2303.15343. arXiv:2303.15343 [cs].

[41] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021. URL http://arxiv.org/abs/2106.09685.

[42] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna M. Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature genetics*, 45(10):1113–1120, October 2013. ISSN 1061-4036. doi: 10.1038/ng.2764. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3919969/.

[43] Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. HISTAI: An Open-Source, Large-Scale Whole Slide Image Dataset for Computational Pathology, May 2025. URL http://arxiv.org/abs/2505.12120. arXiv:2505.12120 [eess].

[44] Varun Belagali, Saarthak Kapse, Pierre Marza, Srijan Das, Zilinghan Li, Sofiène Boutaj, Pushpak Pati, Srikar Yellapragada, Tarak Nath Nandi, Ravi K Madduri, et al. Ticon: A slide-level tile contextualizer for histopathology representation learning. *arXiv preprint arXiv:2512.21331*, 2025.

[45] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

A Appendix

A.1 Supervised Implementation

We follow the standard training recipe for MIL models, as described in [11, 35]. The experimental setup utilized the AdamW optimizer with a base learning rate of 1×10^{-4} and a cosine annealing schedule. To optimize computational efficiency, we employed PyTorch's native mixed-precision training. For benchmarks providing a validation split, models were trained for a maximum of 20 epochs, incorporating an early stopping mechanism with a 5-epoch patience window and a 10-epoch minimum threshold. The optimization objective was cross-entropy loss, paired with a random class-weighted sampling strategy and a batch size of 1. To prevent overfitting, we applied a weight decay of 1×10^{-5} alongside a dropout rate of 0.25 within all feedforward layers and 0.1 for features extracted from the frozen encoder. Training and evaluation is performed on three NVIDIA RTX 3090 GPUs. Training for three epochs on the full training set was completed in 31 hours.

A.2 Dataset Descriptions

Mass-340K Training Dataset. We use the Mass-340K dataset used to pretrain TITAN, comprising 335,645 WSIs across 20 organs from Mass General Brigham institutional and consultation collections and the GTEx consortium [6]. The corpus includes H&E, IHC, and special stains; associated diagnostic text includes clinical reports and GTEx pathology notes. We refer to TITAN for the original dataset curation and governance details.

Our adaptation corpus contains 335,645 slide–text pairs and excludes TCGA, CAMELYON, HistAI/SPIDER, and CATCH slides and reports. The separate public-data adaptation experiment uses TCGA and is evaluated on SPIDER and CATCH. No additional manual slide or report annotation was performed for LIMIT. Internal data were de-identified before analysis. We normalize the associated diagnostic text using a locally hosted Qwen3-14B model and the prompt in Table A1.

Table A1: Prompt used to normalize pathology reports in the internal training dataset. The placeholder `{report}` denotes the raw pathology report text.

Role	Prompt content
System	You are a pathology text normalizer. Extract the diagnostic information from the text. Remove extraneous details not visible from histology, such as laterality, staining, immunohistochemistry, marker names, and positivity/negativity of genetic testing. Return the cleaned diagnosis text.
User	Clean this pathology report. Report: `{report}` Example: Input: Metastatic poorly-differentiated adenocarcinoma with giant cell features identified in the liver, positive for CK7, GATA3, and focal p40, with PD-L1 expression in tumor cells. Output: Metastatic poorly-differentiated adenocarcinoma with giant cell features

TCGA-OT. The TCGA-OT dataset [6] comprises 11,186 diagnostic H&E formalin-fixed paraffin-embedded (FFPE) WSIs of primary tumors drawn from The Cancer Genome Atlas (TCGA) [42], a public consortium that aggregated tissue specimens contributed by tissue source sites at U.S. and Canadian cancer centers under the auspices of the National Cancer Institute and National Human Genome Research Institute. Each slide was assigned to one of 46 OncoTree disease subtypes by mapping the original TCGA clinical diagnosis (assigned by the contributing institution's pathologists at the time of sign-out) to the standardized OncoTree ontology, with the constraint that each retained class contain at least 50 samples using the splits provided by TITAN [6]. IHC preparations, frozen

sections, slides missing magnification metadata, samples with insufficient tumor, and slides from recurrent or metastatic growths were excluded. During quality control, we found that the OncoTree codes for melanoma (MEL) and cutaneous melanoma (SKCM) were not applied consistently across TCGA: both comprise predominantly cutaneous melanomas, and the assignment appears to reflect whether a subtype was recorded at sign-out rather than a histological distinction. We therefore merged the two into a single melanoma class, yielding 45 classes. The dataset is partitioned into 8,226 training, 1,612 validation, and 1,348 test slides.

Rather than a 45-way tissue-agnostic classification, we formulate this as an organ-specific multiple-choice task, mirroring clinical workflows where tissue origin is known *a priori*. For each WSI, valid diagnostic options are restricted to its corresponding organ. To account for 11 of the 28 organ codes containing only a single diagnostic label, we group bladder (n=457), prostate (n=449), cervix (n=229), and the remaining cases (n=409) to a single category. For reporting concision, we report the 7 organs with the highest volume of slides, and group the remaining into an "Other" category, reported as macro-averaged balanced accuracy.

SlideChat-HistAI. We use a HistAI-derived multiple-choice subset of the SlideChat evaluation suite, comprising histological subtyping over 1,396 WSIs [9]. The underlying slides are drawn from the HistAI open-source whole-slide image dataset [43], an aggregation of de-identified diagnostic WSIs scanned predominantly on Leica Aperio GT450 and AT2 systems and curated by HistAI, accompanied by structured clinical metadata (diagnosis, demographics, standardized diagnostic codes). While originally designed for LLM-based question answering, the same multiple choice options are provided for all slides within an organ. Consequently, we treat this as a three-way zero-shot classification task and report balanced accuracy within each organ.

SPIDER. The Supervised Pathology Image-DEscription Repository (SPIDER) [29] is a patch-level multi-organ benchmark released by HistAI, with WSIs sourced from the broader HistAI WSI dataset [43]. We evaluate on the four publicly released organ subsets: skin (159,854 patches, 24 classes), breast (92,892 patches, 18 classes), colorectal (77,182 patches, 13 classes), and thorax (78,307 patches, 14 classes). Annotations were generated by board-certified pathologists using HistAI's web-based CELLDX annotation platform: for each candidate slide, pathologists delineated regions of the target morphology with polygon and brush tools, and the annotated regions were tessellated into non-overlapping 224×224 pixel patches at $20\times$ magnification. Each released patch is paired with surrounding context patches and a label verified by an expert pathologist. We report both fine-grained classification (SPIDER-F) using the official class labels and a unified 7-class mapping (SPIDER-C: Tumor, Stroma, Inflammation, Necrosis, Epithelium, Vascular, Fat), with additional categories, Nerve and Skin adnexa (Skin) and Mucin (Colorectal), defined in consultation with an in-house pathologist.

CAMELYON-16 and CAMELYON+. CAMELYON-16 [30] consists of 400 H&E-stained WSIs of sentinel lymph node sections from breast cancer patients, collected at Radboud University Medical Center (Nijmegen, the Netherlands) and the University Medical Center Utrecht (Utrecht, the Netherlands). The original ground truth was produced by pathologists at the contributing institutions, who delineated metastatic regions on each tumor-positive slide as polygon contours stored in XML and binary mask formats; non-tumor slides were verified as negative. Because the original CAMELYON-16 annotations are known to contain errors and incomplete pixel-level coverage, we use the CAMELYON+ re-annotation by Ling *et al.* [31], which jointly reprocesses the CAMELYON-16 and CAMELYON-17 collections: a panel of expert pathologists removed low-quality slides, corrected erroneous slide-level labels, and provided dense pixel-level tumor annotations including for slides previously lacking them. From the CAMELYON+ pixel-level masks we derive patch-level labels of metastasis vs. non-cancerous lymph node tissue, yielding 206,800 patches across 48 slides (71,564 with metastasis).

CATCH. The Pan-Tumor Canine Cutaneous Cancer Histology (CATCH) dataset [32] consists of 350 H&E-stained WSIs of cutaneous tissue biopsies from 282 canine patients, retrospectively selected from the biopsy archive of the Institute for Veterinary Pathology, Freie Universität Berlin. Slides were stained with routine hematoxylin and eosin and digitized on Leica linear scanning systems at $0.25\,\mu m/$pixel. The dataset includes 12,424 polygon annotations across 13 classes: six non-neoplastic tissue classes (epidermis, dermis, subcutis, bone, cartilage, and a joint inflammation/necrosis class) and seven cutaneous tumor subtypes (melanoma, mast cell tumor, squamous cell carcinoma, peripheral

nerve sheath tumor, plasmacytoma, trichoblastoma, and histiocytoma). Annotations were drawn by veterinary pathologists using the open-source SlideRunner tool, and label quality was validated through inter-rater experiments with three independent pathologists, demonstrating high concordance particularly for tumor regions. We tessellate the annotated WSIs into patches and assign a patch the dominant class label only when at least 99% of the patch area falls within a single annotated class; patches with conflicting labels are excluded. Under-represented tissues (bone, cartilage) are excluded following previous pipelines [44], yielding 11 classes. Final patch counts are reported in Table A2.

Table A2: Patch Counts by Label for CATCH dataset. A patch label was given a label if >99% of the area within that patch consisted of that label. Patches without >99% area containing a single label were not included for analysis (conflicting labels).

Label	**Count**
Dermis	263,813
Epidermis	23,579
Histiocytoma	50,770
Inflamm/Necrosis	32,723
Mast Cell Tumor	166,951
Melanoma	118,835
PNST	193,608
Plasmacytoma	84,548
SCC	63,839
Subcutis	90,682
Trichoblastoma	161,933
Conflicting labels	1,978

A.3 Baseline Model Descriptions

We evaluate CONCH-Z against six published VL pathology baselines spanning patch-level and slide-level granularities. We summarize the data scale and pretraining approach of each model in Table A3, and provide individual descriptions below.

PLIP [13]. Pathology Language-Image Pretraining (PLIP) is a CLIP-style VL model produced by fine-tuning OpenAI's CLIP (ViT-B/32 image encoder, transformer text encoder) on OpenPath, a corpus of 208,414 image-text pairs curated from pathology-related Twitter hashtags (collected between March 2006 and November 2022) and supplemented with pathology subsets of LAION-5B. Captions originate from the natural-language posts and replies of pathologists on social media, with quality filtering applied during curation. Training uses the standard symmetric image-text contrastive (InfoNCE) objective.

QuiltNet [14]. QuiltNet is a CLIP fine-tuned on Quilt-1M, a corpus of approximately one million pathology image-text pairs. The Quilt subset (768K–802K pairs depending on release) is extracted from 1,087 hours of educational pathology videos across 4,504 YouTube videos using a pipeline that combines automatic speech recognition, GPT-3.5-based ASR error correction, and UMLS-grounded text post-processing to align narrated commentary with extracted video frames. The remainder of Quilt-1M is drawn from Twitter, research papers, and the open web (including OpenPath and ARCH). Training follows the standard CLIP contrastive objective. We use the QuiltNet-b16 variant for this study.

MUSK [4]. MUSK (Multimodal transformer with Unified maSK modeling), developed by Stanford and Harvard, is a BEiT-3-based VL model with shared self-attention blocks and modality-specific feed-forward experts. The first pretraining stage performs masked modeling independently on 50 million pathology image patches (extracted from approximately 33,000 WSIs spanning 11,577 patients across 33 tumor types) and on 1 billion pathology-related text tokens, without requiring image-text pairing. The second stage performs CLIP-style contrastive alignment on the 1M paired Quilt-1M dataset [14], augmented with an auxiliary masked language modeling loss.

MI-Zero [5]. MI-Zero is a multiple instance learning framework for zero-shot WSI classification built on a CLIP-style patch encoder. The released encoders use CTransPath as the vision backbone and either BioClinicalBERT or PubMedBERT as the text encoder, fine-tuned via InfoNCE on a corpus of 33,480 patch-level image-caption pairs assembled from the ARCH dataset (7,562 pairs from educational material and PubMed figure legends) and additional internal data. At inference, slide-level predictions are obtained by computing patch-level cosine similarities to class prompts and aggregating via top-K pooling. We use the BioClinicalBERT variant for this study.

CONCH and CONCHv1.5 [3, 6]. CONCH (CONtrastive learning from Captions for Histopathology) is a CoCa-style VL patch encoder pretrained on 1.17 million human histopathology image-caption pairs curated from the PubMed Central Open Access (PMC OA) dataset and educational sources, with three-stage automated cleaning (histopathology image detection, caption splitting, and image-caption matching). The training objective combines contrastive image-text alignment with an auxiliary captioning loss. CONCHv1.5 upgrades the original CONCH VL patch encoder by integrating a Vision Transformer Large (ViT-L/16) backbone restored from the UNI checkpoint [45] and fine-tuning it under the same CoCa objective on a slightly larger and more diverse 1.26 million image-caption corpus that spans FFPE, frozen, and IHC tissue types.

TITAN [6]. TITAN is a slide-level VL model pretrained on Mass-340K [6]. Its three stages comprise visual self-supervised learning on 335,645 WSIs, alignment with 423,122 synthetic ROI captions, and alignment with 182,862 slide–report pairs drawn from institutional clinical reports and GTEx pathology notes. TITAN operates on CONCHv1.5 patch features.

PRISM [15]. PRISM (Pathology Report and Image Summarization Model) is a slide-level VL model pretrained on an internal corpus of 587,000 WSIs and 195,000 clinical pathology reports from Memorial Sloan Kettering Cancer Center; reports were rewritten using GPT-4 prior to training. PRISM consists of a Perceiver-based slide encoder operating on Virchow V1 patch tile embeddings and a BioGPT-based language decoder, jointly trained with the CoCa objective (contrastive image-text alignment plus autoregressive captioning). The model contains $\sim$558M parameters.

A.4 Ablation Implementation

Sigmoid Locked-Image Tuning (SigLIT)

Sigmoid Locked-Image Tuning (SigLIT). Our weighted SigLIT ablation [40] retains both report-conditioned pooling and the detached inter-report similarity weights used in LIMIT, replacing the InfoNCE objective with a pairwise sigmoid loss. For a batch of K slide–report pairs, define $y_{ij} = 1$ when $i = j$ and $y_{ij} = -1$ otherwise. Each slide–report combination uses its own report-conditioned representation $\mathbf{h}_{W_i}(\mathbf{r}_j)$. We assign positive pairs unit weight and negative pairs the same weights as in Section 3.2:

$$a_{ij} = \begin{cases} 1, & i = j, \\ \mathrm{stopgrad}\left[\dfrac{1 - \mathrm{sim}(\mathbf{r}_i, \mathbf{r}_j)}{2}\right], & i \neq j. \end{cases} \tag{7}$$

The loss averages the weighted binary classification terms over all slide–report combinations:

$$\mathcal{L}_{\mathrm{SigLIT}} = -\frac{1}{K^2} \sum_{i=1}^{K} \sum_{j=1}^{K} a_{ij} \log \sigma\left(y_{ij} \left[\frac{\mathbf{h}_{W_i}(\mathbf{r}_j)^\top \mathbf{r}_j}{\tau} + b\right]\right), \tag{8}$$

where σ is the sigmoid function, τ is a learnable temperature, and b is a learnable bias. Semantically overlapping negative reports receive smaller weights. Gradients flow through the alignment scores and report-conditioned pooling, but not through a_{ij}; the vision encoder remains frozen.

Training Sizes: To obtain datasets at 150k and 50k samples, we perform stratified sampling without replacement to preserve the distribution of organs on the full pretraining set (**Figure A3**). We evaluate the influence of inference-time ablations (mean pool, random pool) at the 150k and 50k scale using the models trained with the LIMIT approach.

Table A3: Summary of baseline vision-language pathology models. "Vision corpus" reports the scale of unique image data (image-text pairs, patches, or WSIs) used for VL alignment. "Source institutions" lists the primary clinical or public sources of training data; public corpora are denoted in italics. Citations refer to the model release papers.

Model	Architecture	Vision corpus	Source institution(s)	Objective	Granularity
PLIP [13]	ViT-B/32 + Transformer	208K pairs	*Twitter (#path), LAION-5B*	CLIP InfoNCE	Patch
QuiltNet [14]	ViT-B/16 + Transformer	1M pairs	*YouTube, Twitter, web, ARCH*	CLIP InfoNCE	Patch
MI-Zero [5]	CTransPath + BioClinicalBERT	33K pairs	*ARCH* + internal	InfoNCE + MIL pool	Patch / WSI
MUSK [4]	BEiT-3 (multiway transformer)	50M patches (∼33K WSIs) + 1M pairs	Stanford internal; *Quilt-1M*	Masked modeling + CLIP	Patch
CONCHv1.5 [3, 6]	ViT (CoCa, UNI-init)	1.26M pairs	*PMC OA, EDU*	Contrastive + captioning	Patch
TITAN [6]	ViT slide encoder (iBOT + CoCa)	335K WSIs + 423K ROI-captions	BWH, MGH; *GTEx*	SSL + contrastive + captioning	WSI
PRISM [15]	Perceiver + BioGPT	587K WSIs (Virchow tiles)	MSKCC	CoCa (contrastive + captioning)	WSI
CONCH-Z (ours)	CONCHv1.5 (frozen) + adapted text encoder	335,645 slide–text pairs	BWH, MGH; *GTEx*	Difficulty-weighted contrastive (LIMIT)	Patch / WSI

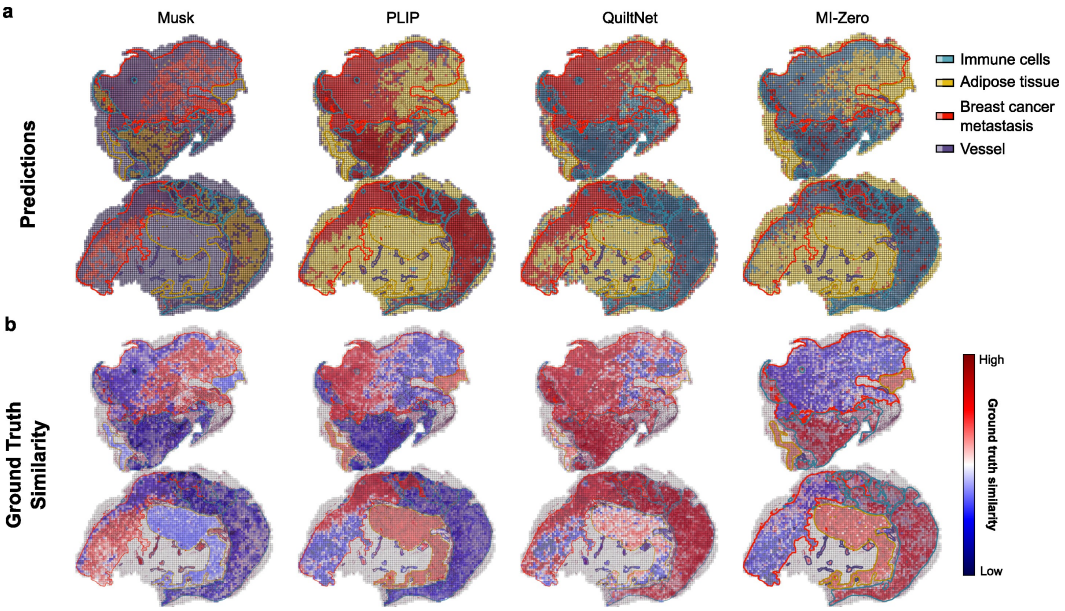

Figure A1: **Zero-shot breast cancer metastasis in lymph nodes for various VL models**. **a)** Patch-level predictions according to the class embedding with the highest cosine similarity to each patch embedding. Annotated regions according to an in-house pathologist are overlayed. **b)** Direct similarity scores between each class embedding and the patch embedding. More red indicates higher similarity. Scores are normalized within each model. Among the baseline patch models, QuiltNet exhibits the highest degree of overlap with the annotated regions.

Table A4: Terms and Prompt Templates for SPIDER-C. Synonyms are used as interchangeable terms, and templates define text prompt structures where {term} is replaced by each synonym.

Category	Content
Synonyms	Tumor; Cancer; Malignancy; Carcinoma; Neoplasm
Templates	{term} an image of {term} {term} cells

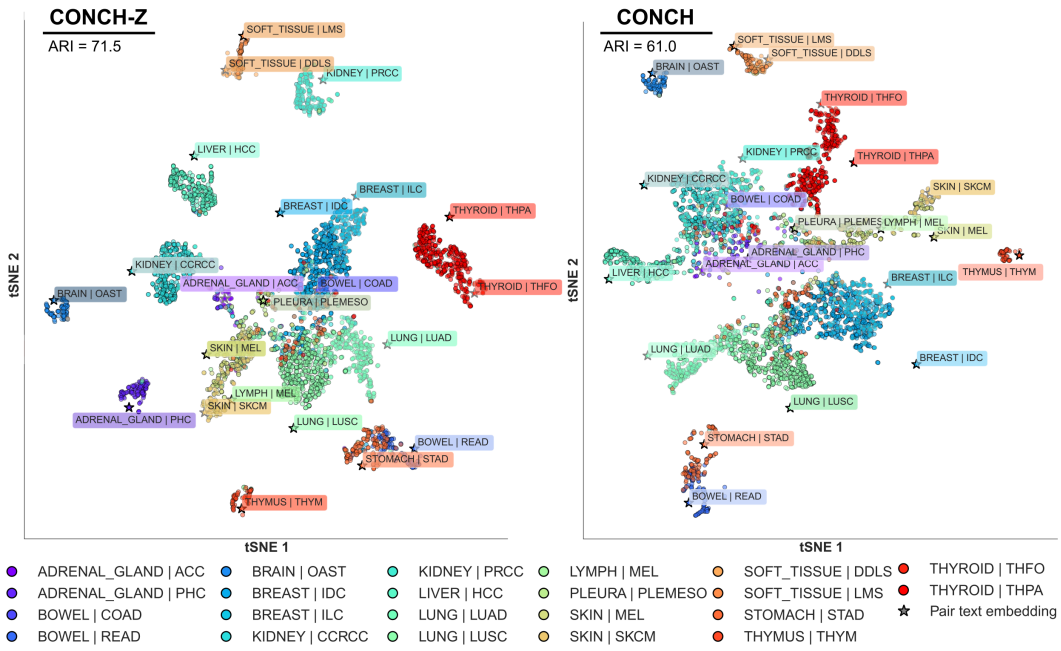

Figure A2: **Alignment of text embeddings with slide representations on TCGA-OT**. T-SNE on top 10 components of canonical correlation analysis of cancer subtypes and slide-level features. Slide-level features were obtained by weighted average with query text embedding "cancer". Textual embeddings of CONCH-Z exhibit higher overlap with image embeddings than CONCHv1.5 text embeddings. For instance, Adjusted rand index (ARI) indicates substantially better alignment for CONCH-Z (ARI=71.5) compared to CONCHv1.5 (ARI=61.0).

A.5 Additional Pretraining-Corpus Characterization

The corpus contains 148,279 unique normalized diagnostic phrases and 689 unique organ strings. These strings are raw textual variants rather than 689 distinct anatomical classes; the main corpus description uses 20 organ groups. The normalized text retains compositional diagnostic descriptors, such as differentiation and architectural features, but is not the full original report.

Table A5: **Pretraining-corpus composition.** Clinical groups were assigned using an LLM; percentages describe these automatically assigned categories.

Category	Group	Percentage
Clinical group	Malignant/neoplastic	69.2
	Non-neoplastic	23.6
	Premalignant/benign	4.1
	Unresolved	3.1
Acquisition metadata	Leica	35.0
	Motic	32.3
	Philips	9.9
	3D Histech	9.1
	GTEx source; vendor unspecified	7.4
	Hamamatsu	6.2
	Roche/Zeiss	< 1
Native resolution	0.24–0.26 μm/pixel	52.9
	0.49–0.52 μm/pixel	39.9
	0.12–0.23 μm/pixel	7.1

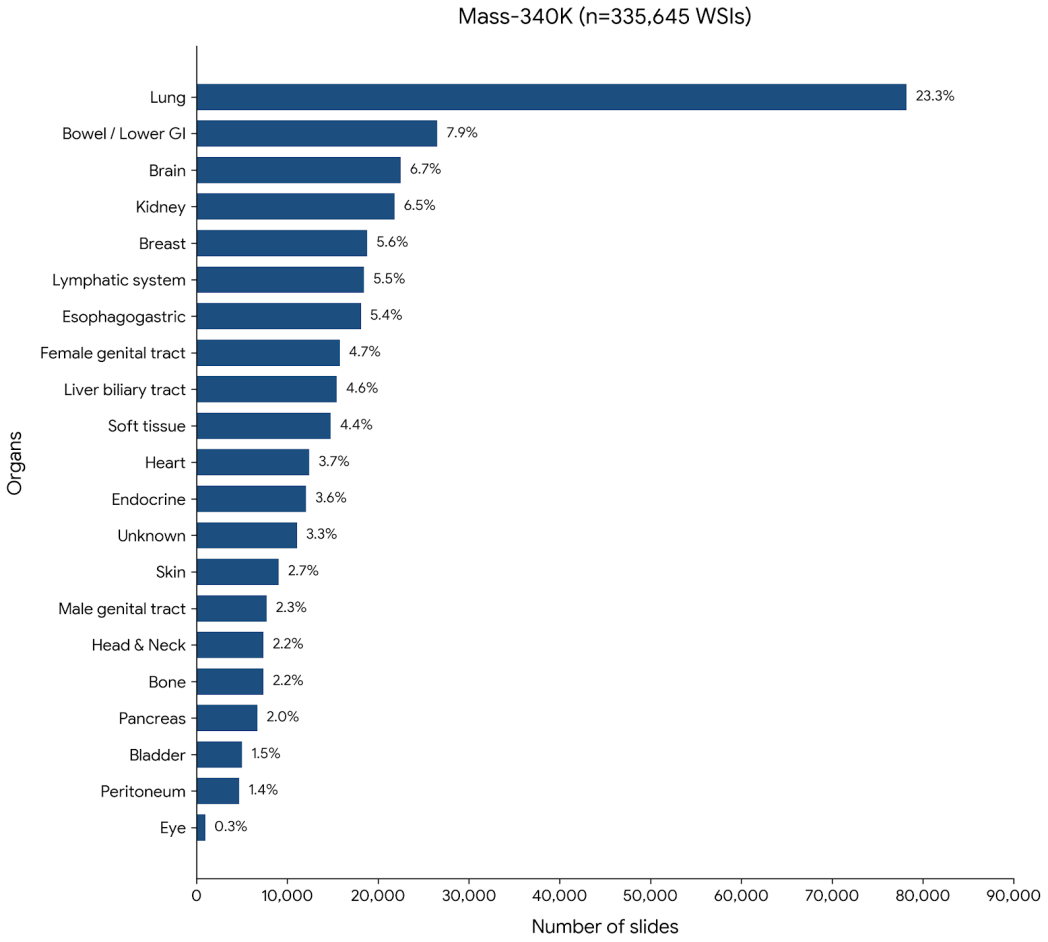

Figure A3: Training data distribution by organ.

A.6 Diagnostic-Term Exposure and Zero-Shot Interpretation

We examined report-level mentions of downstream diagnostic labels and synonyms in the internal adaptation corpus. Each report was lowercased, hyphens and slashes were replaced by spaces, brackets and separators (, : ;) were removed, and whitespace was collapsed. Class terms were then matched as whole words, so a term inside a longer word does not count. Negations (e.g., "negative for melanoma") were not excluded, so the counts are an upper bound on diagnostic mentions. Table A6 summarizes diagnostic-term exposure and its association with per-class performance gains. The correlation is computed between per-class exposure (the number of reports mentioning the class) and per-class gain (accuracy of CONCH-Z minus CONCH).

Table A6: Reported lexical exposure analysis. “No match” is the percentage of reports without a matched diagnostic term. Correlations relate diagnostic exposure to per-class performance gain.

Benchmark	No match (%)	Spearman ρ	p
CATCH-F	95.0	0.414	0.36
HistAI	99.7	0.54	0.21
SPIDER-F	95.6	-0.05	0.81
TCGA-OT (45 classes)	90.3	0.07	0.72

Across these analyses, we did not detect a statistically significant association between reported diagnostic exposure and performance gain.

A.7 Cross-Backbone Adaptation

Table A7: **Cross-backbone adaptation.** Change in balanced accuracy (percentage points) after LIMIT adaptation on 25,000 Mass-340K slide–text pairs. Explicit signs indicate improvement or decline relative to each model's unadapted baseline. Overall is the equal-weight mean across the 17 tasks.

Dataset	Task	PLIP	QuiltNet	MUSK
SPIDER-C	Breast	+3.96	+4.31	+5.19
	Colorectal	+0.17	+1.94	+3.31
	Thorax	+6.81	+2.33	+3.92
	Skin	+1.38	+0.50	+0.24
SPIDER-F	Breast	+4.13	+1.79	+1.50
	Colorectal	+2.86	+3.72	+2.40
	Thorax	+1.22	−1.32	+4.33
	Skin	+4.42	+1.96	+0.51
CATCH	Skin-C	+2.50	+4.00	+1.80
	Skin-F	+1.30	+6.60	+5.60
HistAI	Breast	+7.30	+5.20	+3.10
	Colorectal	+3.60	+4.70	+8.40
	Skin	+2.80	+4.80	+1.10
TCGA-OT	Breast	+1.00	+4.50	+4.30
	Lung	+7.10	−0.30	+6.80
	Brain	+2.70	+4.80	+1.30
	Other	+4.60	+2.70	+4.20
Overall		**+3.40**	**+3.07**	**+3.41**

A.8 Adaptation Using Public TCGA Reports

We applied LIMIT to CONCHv1.5 using 10,108 public TCGA WSIs and their associated pathology reports. The overall changes in balanced accuracy are +6.60 percentage points after public-data adaptation and +7.00 percentage points after internal-data adaptation, relative to the corresponding unadapted CONCHv1.5 baseline. Both settings improve all ten evaluated tasks.

Table A8: **Adaptation with public and institutional training data.** Change in balanced accuracy (percentage points) relative to the unadapted CONCHv1.5 baseline. LIMIT is trained on TCGA (10,108 WSIs) or Mass-340K (25,000 slide–text pairs).

Dataset	Task	TCGA 10k	Mass-340K 25k
SPIDER-F	Skin	+1.10	+4.50
	Colorectal	+4.40	+3.70
	Breast	+2.20	+0.30
	Thorax	+3.40	+3.60
	Average	**+2.80**	**+3.00**
SPIDER-C	Skin	+1.30	+8.60
	Colorectal	+13.10	+10.60
	Breast	+15.60	+20.70
	Thorax	+15.50	+9.60
	Average	**+11.40**	**+12.40**
CATCH	Skin-F	+5.00	+5.70
	Skin-C	+4.30	+2.30
	Average	**+4.65**	**+4.00**
Overall		**+6.60**	**+7.00**

A.9 Visual Adaptation: Performance and Compute

LIMIT-V first ranks patches using frozen features and then dynamically re-encodes selected patches for vision-side gradient updates. The following timing measurements were reported for a single training sample, averaged over 100 samples.

Table A9: **Training cost of locked-image and visual adaptation.** Memory is peak GPU memory in GB; forward/backward times are seconds. OOM denotes out of memory under the evaluated setup.

	Memory (GB)		Forward (s)		Backward (s)	
k	LIMIT	LIMIT-V	LIMIT	LIMIT-V	LIMIT	LIMIT-V
8	4.3	14.4	.126	.577	.057	.291
16	4.3	20.2	.127	.802	.059	.530
32	4.3	31.9	.133	1.168	.060	.753
64	4.3	55.3	.133	4.687	.063	1.435
128	4.3	OOM	.138	OOM	.064	OOM
All	4.8	OOM	.184	OOM	.082	OOM

Table A10: **Visual fine-tuning performance.** Change in SPIDER balanced accuracy (percentage points) relative to the CONCHv1.5 baseline, after fine-tuning on 25,000 internal pairs. Deltas are computed from the reported task-level scores; averages weight the four organs equally. LIMIT-V uses $k = 64$ selected patches.

Dataset	**Task**	**LIMIT**	**LIMIT-V**
SPIDER-F	Skin	$+4.50$	$+1.10$
	Colorectal	$+3.70$	-0.20
	Breast	$+0.30$	$+7.10$
	Thorax	$+3.60$	$+3.40$
	Average	$\mathbf{+3.03}$	$\mathbf{+2.85}$
SPIDER-C	Skin	$+8.60$	$+0.40$
	Colorectal	$+10.60$	$+13.10$
	Breast	$+20.70$	$+22.30$
	Thorax	$+9.60$	$+9.40$
	Average	$\mathbf{+12.38}$	$\mathbf{+11.30}$

A.10 Text-Encoder Initialization

All conditions retain a frozen pretrained CONCHv1.5 vision encoder. "Random" uses a frozen randomly initialized text encoder; "Random-trained" adapts that text encoder. In "Pretrained queries," a separate frozen CONCHv1.5 text encoder supplies the pooling queries while a randomly initialized trainable text encoder supplies the report embeddings used for alignment.

A.11 Candidate-Count Scaling

We benchmark query-guided pooling on a slide with 10,000 patches and 10–300 candidate labels on one Nvidia A100. Table A12 presents the individual model measurements and their arithmetic mean. The measurements apply to the aggregation step, rather than end-to-end slide processing. Mean pooling latency increases from 7.32 ms at 10 candidates to 8.02 ms at 300 candidates, an endpoint slope of approximately 0.0024 ms per additional candidate. These measurements suggest modest incremental pooling cost over the tested label range; the operation still scales with patch count, candidate count, and embedding dimension.

Table A11: **Text-initialization ablation.** Change in balanced accuracy (percentage points) relative to the frozen randomly initialized text encoder (Random). CONCH-Z@25k uses a pretrained text encoder followed by LIMIT adaptation on 25,000 internal pairs. Deltas are computed from the reported task-level scores; averages weight the four organs equally.

Dataset	Task	Random-trained	Pretrained queries	CONCH-Z @25k
SPIDER-F	Skin	+3.30	+17.60	+38.60
	Colorectal	+7.10	+15.20	+31.80
	Breast	+3.20	+22.60	+22.20
	Thorax	+10.00	+11.40	+38.20
	Average	**+5.90**	**+16.70**	**+32.70**
SPIDER-C	Skin	+7.00	+10.30	+38.10
	Colorectal	+5.30	+22.60	+29.00
	Breast	+6.30	+32.60	+49.50
	Thorax	+6.70	+18.20	+19.00
	Average	**+6.33**	**+20.93**	**+33.90**

Table A12: Pooling latency in milliseconds versus number of candidate labels. The mean weights the five models equally.

Model	10	25	50	75	100	200	300
PLIP	11.1	11.2	11.2	11.2	11.2	11.5	11.4
QuiltNet	3.1	3.2	3.2	3.2	4.0	3.6	3.9
CONCH	13.2	13.3	13.4	13.5	13.5	13.9	14.1
MUSK	6.0	6.0	6.0	6.1	6.1	6.6	6.9
MI-Zero	3.2	3.0	3.2	3.3	3.4	3.6	3.8
Mean	7.32	7.34	7.40	7.46	7.64	7.84	8.02

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions: LIMIT, locked-image tuning of the CONCH v1.5 text encoder using 335,645 clinical reports, and evaluation across patch-level and slide-level zero-shot benchmarks. The reported results in Sections 4.1–4.5 support these claims, and the limitations section narrows the scope of the conclusions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

Justification: The paper explicitly discusses limitations, including the absence of tunable image-side parameters, possible difficulty with holistic slide-level descriptors such as grade or tumor cellularity, the diagnosis-focused evaluation scope, and dependence on the frozen CONCH v1.5 visual backbone.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results, theorems, or proofs. The mathematical content consists of the proposed report-guided attention and contrastive optimization formulation.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The paper provides the model formulation, training objective, inference procedure, datasets, baselines, hyperparameters, report preprocessing prompt, evaluation settings, and compute setup needed to understand and reproduce the experimental protocol. Although the internal clinical WSI/report corpus cannot be publicly released, we will release CONCH-Z checkpoints and evaluation scripts upon acceptance, providing a practical path for researchers to reproduce or verify the main benchmark results using the public evaluation datasets.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [No]

Justification:The internal clinical WSI/report corpus used for training cannot be openly released due to patient privacy and institutional data-use restrictions. Upon acceptance, we will release CONCH-Z model checkpoints and evaluation scripts for the public benchmarks, but these assets are not yet openly available at submission time.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper specifies the main experimental settings, including task formulations, datasets, baselines, patch size, negative report sampling, number of negatives, optimizer, learning rate, weight decay, validation split, training epochs, report preprocessing, and supervised MIL training details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The main result tables report standard deviations, with several tables specifying that uncertainty was estimated using 1,000 bootstrap trials. These error estimates are provided for the patch-level, slide-level, and supervised baseline comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper states that training and evaluation were performed on three NVIDIA RTX 3090 GPU, and training for three epochs on the full training set was completed in 31 hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The study uses de-identified retrospective pathology data under IRB approval and evaluates computational pathology models without claiming deployment as a clinical diagnostic system.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The appendix includes a Broader Impacts subsection discussing potential benefits for computational pathology as well as risks from incorrect predictions, distribution shift, bias, privacy considerations, and the need for prospective clinical validation before deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The current work does not release high-risk dataset or broadly deployable model requiring special misuse safeguards. The internal clinical reports and WSIs are described as de-identified and are not stated to be publicly released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [No]

    Justification: While the paper cites existing datasets, models, and tools, including TCGA, SlideChat-HistAI, SPIDER, CAMELYON/CAMELYON+, CATCH, PLIP, QuiltNet, MUSK, MI-Zero, CONCH v1.5, TITAN, PRISM, Trident, and Qwen3, it does not list their licenses or terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets`

has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The paper introduces CONCH-Z as a new model asset, but the model checkpoints and evaluation scripts are not publicly available at submission time. Upon acceptance, we will release the model checkpoints and evaluation scripts with documentation describing intended use, limitations, access instructions, and licensing; the internal clinical WSI/report training corpus will not be released due to institutional restrictions.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [No]

    Justification: The paper does not involve crowdsourcing, prospective participant interaction, participant instructions, screenshots, or compensation. The human clinical data are retrospective and de-identified.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [Yes]

    Justification: The study used retrospective, de-identified pathology slides and reports under institutional review board approval. Institution-specific IRB details are omitted from the anonymous submission and will be provided in the non-anonymous final version if accepted.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: The paper describes using a locally hosted Qwen3-14B model to remove residual identifying information and filter report content unrelated to histology, and it provides the report-normalization prompt in Table A1.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.