Scaling Whole-Slide Pathology Foundation Model Pretraining with Billions Off-the-Shelf Tokens

Honglin Li[1,2], Zhongyi Shui[1,2], Chenglu Zhu[1,3⋆], and Lin Yang[1,3⋆]

[1] School of Engineering, Westlake University, Hangzhou, 310030, China
[2] College of Computer Science and Technology, Zhejiang University, Hangzhou, 310058, China
[3] Institute of Advanced Technology, Westlake Institute for Advanced Study, Hangzhou, 310024, China.
{lihonglin,zhuchenglu,yanglin}@westlake.edu.cn

Abstract. Pathology whole slide image (WSI) analysis is crucial for disease diagnosis and understanding. While pathology foundation models have advanced WSI analysis, their scalability remains limited by the large resolution of individual WSIs and the small number of available slides. Existing methods often use the [CLS] token from tile-level ViTs for efficiency, but this design overlooks spatially informative tokens that encode fine-grained pathological patterns. Incorporating all spatial tokens could enrich representation but is computationally prohibitive, revealing a fundamental trade-off between efficiency and representational richness. We observe high redundancy among spatial tokens, suggesting that selectively retaining the most informative ones can better balance these goals. To address this, we propose TokRet, a token retention and compression module that preserves the most informative spatial tokens while eliminating redundancy. Built upon TokRet, we develop PathTokScale, a slide-level self-supervised pretraining framework that leverages efficient long-contextual transformer modeling to scale training to the billion-token level—over 20× larger than previous WSI studies—without requiring massive WSI collections. Experiments show that PathTokScale consistently outperforms existing pathology foundation models across diagnostic, prognostic, and biomarker prediction tasks. Remarkably, using only 10% of the number of WSIs as used by TITAN, it achieves comparable performance through efficient token scaling, paving the way for scalable, high-performing pathology foundation models.

Keywords: Computational Pathology · Whole Slide Image · Foundation Model

1 Introduction

Pathology whole slide image (WSI) analysis is central to computational approaches for disease diagnosis, prognosis, biomarker discovery and gene expression [6, 12, 13, 16, 22, 24, 58, 59, 70]. WSIs capture rich spatial and morphological

⋆ Corresponding authors

information crucial for clinical decision-making, but their **Giga-pixel high resolution and limited number of annotated slides** pose major challenges for developing scalable and generalizable models [1, 31, 71]. To mitigate reliance on exhaustive annotations, weakly-supervised [7, 10, 46] and self-supervised learning (SSL) [10,11,75] paradigms have emerged as promising directions in computational pathology.

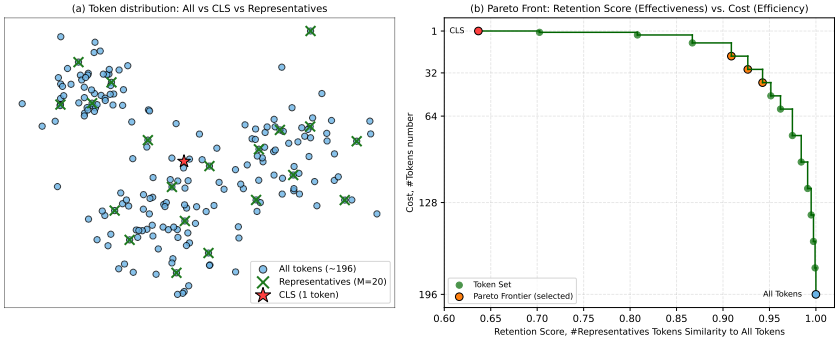


Fig. 1: Visualization of representative token selection and Pareto efficiency trade-off. (a) Spatial distribution of all tokens extracted from a pathology patch ViT embeddings, the [CLS] token, and a selected subset of representative tokens ("Ours"). The plot illustrates how our method selects a small number of tokens while maintaining coverage of the full embedding distribution. (b) Pareto front showing the trade-off between cost (tokens number) and retention score (cosine similarity to all tokens). The highlighted range indicates the Pareto-optimal region where a small token subset achieves a good balance between efficiency and effectiveness. [CLS] and all-token sets are marked for reference.

Recently, foundation models (FMs) for computational pathology have accelerated progress by learning transferable representations across diverse downstream tasks. Self-supervised FMs [11,21,50,69] and vision-language (VL) FMs [18,45,54] have demonstrated the ability to extract semantically meaningful features from both tile- and slide-level images. However, VL-FMs are constrained by the scarcity of large-scale image-text paired datasets, limiting their generalization across histological domains. Vision-only FMs trained with SSL benefit from broader data accessibility.

These methods typically extract tile-level features using pretrained ViTs and aggregate them into slide-level representations via multiple instance learning (MIL) [17, 20, 28, 40, 58]. For efficiency, most approaches adopt the task-agnostic [CLS] token as a global tile representation [21, 45], which **overlooks rich spatial information critical for modeling fine-grained pathological variations**. Recent studies [11, 89] show that increasing ViT size provides only marginal gains, revealing a representational bottleneck inherent to [CLS]-only modeling. Attempts to combine [CLS] with pooled spatial tokens [89] yield limited improvements, indicating the importance of preserving informative spatial token representations.

Scaling to a larger number of spatial tokens is particularly beneficial for pathology slide-level analysis, where diagnostic cues are often subtle and spatially distributed. Capturing these fine-grained variations demands broader token coverage. Moreover, expanding model capacity along the token dimension is consistent with the general *scaling law* of large models [26, 30], which predicts improved performance with increased effective token counts.

However, directly utilizing all spatial tokens is computationally prohibitive given the enormous resolution of WSIs because of the quadratic complexity of Transformers [76, 86], leading to an inherent trade-off between efficiency and representational richness. Although recent works attempt to compress these tokens in the feature dimension [36] using auto-encoders [55, 66], such approaches require extra training and often suffer from semantic loss and still yield prohibitively long sequences for Transformer modeling [82, 85]. Interestingly, we observe substantial redundancy among tokens within local tiles or regions (Figure 1b), indicating that **selective token retention** can preserve informative representations while significantly reducing computational cost. This observation resonates with advances in token pruning for natural images [2, 4, 9, 57], though existing approaches typically rely on attention-based heuristics or require costly fine-tuning, which hinders their implementation to pathology WSIs.

To address these challenges, we propose **TokRet**, a plug-and-play selective token retention and compression module that preserves the most informative spatial tokens within each tile-level ViT while eliminating redundancy. TokRet implements a greedy selection strategy by iteratively choosing tokens that are most dissimilar to the already selected ones (covering most embedding distribution shown in Figure 1a), enabling effective reduction of redundancy without requiring additional training. Building upon TokRet, we develop **PathTokScale**, a slide-level self-supervised pretraining framework leveraging efficient long-contextual transformers, enabling billion-token scale training—about 20×larger than prior WSI studies [70, 75], which is off-the-shell from current data-source [49, 65], without requiring massive slide collections.

Our key contributions are summarized as follows:

- We identify a critical scalability bottleneck in existing FMs arising from over-reliance on [CLS] token representations and propose a solution that preserves rich spatial token information.
- We introduce TokRet, a novel token retention and compression framework for WSI analysis that selects the most informative spatial tokens while minimizing redundancy.
- We develop PathTokScale, a slide-level pretraining framework that scales model training to the billion-token level using only off-the-shell public WSI data-resources. Extensive experiments on diagnosis, prognosis, and biomarker prediction tasks demonstrate consistent improvements over [CLS] approaches, establishing new state-of-the-art performance in computational pathology.

By balancing efficiency and representational richness, TokRet and PathTokScale offer a scalable path forward for high-performing foundation models in whole slide image analysis.

2 Related Work

2.1 Pathological Whole Slide Image Analysis

Whole Slide Images (WSIs) contain rich spatial and morphological information essential for pathology [7, 46, 52], but detailed cell-level annotations are labor-intensive, limiting large-scale supervised learning [10]. Weakly- and self-supervised learning have thus become key paradigms for scalable WSI analysis.

Computational Pathology Foundation Models. Early WSI MIL methods rely on tile-level features from pretrained models [12, 17, 28, 46, 47, 80]. Foundation models (FMs) extend this to tile- and slide-level analysis [11, 18, 21, 45, 69, 75], leveraging self-supervised learning (SSL) [8, 14, 15, 50] or image-text multi-modal learning [18, 23, 41, 45, 54] to improve downstream tasks like cancer subtyping and survival prediction.

Fine-tuning tile encoders [38, 63, 81] on high-resolution WSIs or relying all spatial tokens from tiles [36, 52] is computationally intensive, so most methods use pretrained [CLS] tokens as global representations [11, 21, 45], which omits spatial details. Scaling ViTs to larger sizes (e.g., UNI-2 [11]) gives only marginal gains, and combining [CLS] with pooled spatial tokens (e.g., Virchow-2 [89]) yields limited improvement, highlighting a key scalability bottleneck in capturing fine-grained WSI features.

Slide-level FMs Pretraining. Early WSI pre-training methods such as HIPT [10] and Giga-SSL [34] attempted to train on cropped regions extracted from whole slide images (WSIs) and incorporated hierarchical architectures. However, limited data size and suboptimal performance at the tile level constrained their overall effectiveness. This paradigm has also been extended to multimodal WSI pre-training approaches, such as TANGLE [29]. Recently, Slide-level FMs are scaling to large-scale of data and model size, e.g. GigaPath [75] and TITAN [18] employ 6–12 layer Transformers pretrained on tens to millions of WSIs via MAE [25] or iBOT [87], while other SSL approaches [10, 27, 33] use contrastive learning with slide augmentations. However, their scaling is limited by preprocessed tile features, restricting effective large-scale pretraining for high-resolution WSIs.

2.2 Visual Token Pruning

Visual token pruning techniques reduce inference complexity for large models by selecting or merging tokens. One category relies on attention-based criteria, including PruMerge [57] and FastV [9], which cluster or remove tokens based on attention sparsity or magnitude. However, pruning purely by attention scores is often suboptimal [43], especially at high pruning ratios. Calibration-based methods determine pruning layers or ratios using a validation dataset, such as FitPrune [78] and VTW [43], but require additional dataset-specific tuning, limiting general applicability. Some approaches further rely on fine-tuning to learn compact token representations. For example, TokenPacker [39] trains a projector to compress tokens using large-scale data. While effective, these strategies demand

substantial computation and are less practical for domains like pathology with extremely high-resolution images and diverse models.

While prior work on slide-level FMs has advanced WSI analysis, they still face critical limitations: [CLS]-only representations overlook fine-grained spatial details. To address these challenges, we propose **TokRet**, a greedy token selection and compression module that retains the most informative spatial tokens without additional training. Building upon TokRet, **PathTokScale** leverages efficient long-contextual Transformers to enable billion-token scale pretraining, aligning with scaling laws for large models and capturing richer spatial heterogeneity than previous approaches.

3 Method

3.1 Preliminary

For whole slide image (WSI) modeling, a WSI $\mathbf{X}$ is first divided into N smaller tiles: $\mathbf{X} = [\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N]$. Each tile $\mathbf{r}_i$ is further subdivided into n patches: $\mathbf{r}_i = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m]$, where the commonly used tile size is 224×224, and the patch size is 16×16. The patches within each tile are processed by a ViT based FM, such as UNI [11], which outputs representations for each tile:

$$\mathbf{A} = [[\text{CLS}]; \mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_m] \tag{1}$$

Here, the [CLS] token serves as a summary of the information contained in the remaining patch tokens ($\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_m]$). Directly leveraging these $\mathbf{A}$ is computationally prohibitive for WSI modeling, posing need for efficient strategies to handle the large-scale data (e.g. $N \approx 5k$ and $m = 196$ when UNI [11] is applied to some TCGA slides, resulting in about 1 million tokens for a WSI). Without loss of generality, we define an operation $\mathbf{g}$ applied to $\mathbf{T}$ for improving slide-level modeling efficiency. For example, $\mathbf{g}$ can be seen as extracting the [CLS] token [11, 75] or pooling [69, 89] over the patch tokens $[\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_m]$, generating tile embeddings for slide-level input, which are defined as:

$$\mathbf{Z} = [\mathbf{g}(\mathbf{T}_1), \mathbf{g}(\mathbf{T}_2), \ldots, \mathbf{g}(\mathbf{T}_N)] \tag{2}$$

These embeddings are then aggregated to predict the slide-level outcome:

$$\hat{\mathbf{Y}} = \mathbf{f}(\mathbf{Z}; \boldsymbol{\theta}), \tag{3}$$

where $\mathbf{f}(\boldsymbol{\theta})$ is typically implemented using an attention based MIL [28] or a Transformer [58].

3.2 Problem Formulation

Instead of relying solely on the [CLS] token—an efficient but often suboptimal global representation—or retaining all tokens, which ensures completeness but is computationally prohibitive, this work seeks a balanced representation $\mathbf{g}$

that efficiently selects informative tokens for slide-level prediction. Formally, we define the token pruning problem as follows. Given a token set $\mathbf{T}$ with $|\mathbf{T}| = m$, our goal is to select a subset $\tilde{\mathbf{T}}$ of size $\tilde{m}$ ($\tilde{m} < m$) that preserves the key information necessary for accurate prediction. We introduce a mapping function g that transforms the original token set into a reduced one, $\tilde{\mathbf{T}} = \{\tilde{t}_1, \ldots, \tilde{t}_{\tilde{m}}\}$, formulated as:

$$\begin{aligned} \text{Find:} \quad & g : \mathbf{T} \to \tilde{\mathbf{T}} \\ \text{Objective:} \quad & \max_{g} \mathcal{S}\big(\mathbf{T}, \tilde{\mathbf{T}}\big) \\ \text{Subject to:} \quad & |\tilde{\mathbf{T}}| = \tilde{m}, \end{aligned} \tag{4}$$

where $\mathcal{S}(\mathbf{T}, \tilde{\mathbf{T}})$ measures the information retained by the subset $\tilde{\mathbf{T}}$ relative to the original feature set $\mathbf{T}$. To quantify the information retention between the original and reduced token sets, we introduce a cosine similarity–based metric. For each token $\mathbf{t}_i \in \mathbf{T}$, we compute its maximum cosine similarity with all tokens in the reduced set $\tilde{\mathbf{T}}$:

$$s_i = \max_j \cos\left(\mathbf{t}_i, \tilde{\mathbf{t}}_j\right) = \max_j \frac{\mathbf{t}_i \cdot \tilde{\mathbf{t}}_j}{\|\mathbf{t}_i\| \, \|\tilde{\mathbf{t}}_j\|}, \quad i = 1, 2, \ldots, m. \tag{5}$$

Here, $s_i \in [-1, 1]$ measures how well each original token is represented in the reduced subset, with larger values indicating higher alignment. The overall information preservation between the two sets is then measured by averaging these maximum similarities:

$$\mathcal{S} = \frac{1}{m} \sum_{i=1}^{m} s_i = \frac{1}{m} \sum_{i=1}^{m} \max_j \cos\left(\mathbf{t}_i, \tilde{\mathbf{t}}_j\right). \tag{6}$$

The resulting retention score $\mathcal{S}$ serves as a global indicator of similarity between $\mathbf{T}$ and $\tilde{\mathbf{T}}$; a higher $\mathcal{S}$ implies that the pruned token set retains more information from the original features. The Figure 1b provides an illustration on the retention score distribution under our method.

3.3 Token Retention

To achieve efficient yet informative representation, we reformulate the token selection problem as a combinatorial optimization task. The goal is to select a subset of tokens $\tilde{\mathbf{T}}$ that preserves the essential information of the original set $\mathbf{T}$ while minimizing redundancy among the selected tokens. This objective aligns naturally with the Max–Min Diversity Problem (MMDP) [2, 32, 56] in combinatorial optimization [3], which aims to maximize the minimum pairwise distance among chosen elements to ensure maximal diversity:

$$\tilde{\mathbf{T}} = \underset{C \subset \mathbf{T}, |C| = \tilde{m}}{\arg\max} \left[\min_{\gamma, \omega \in C} d(\gamma, \omega) \right], \tag{7}$$

where C is a subset of $\mathbf{T}$ with $\tilde{m}$ tokens, (γ, ω) are arbitrary tokens in C, and $d(\gamma, \omega)$ denotes their cosine distance (Eq. 5). Solving Eq. (7) yields a subset $\tilde{\mathbf{T}}$

Algorithm 1: Greedy Token Retention

Input: $\mathbf{T}$: input token set; $\tilde{m}$: subset size
Output: $\tilde{\mathbf{T}}$: selected subset

Initialize $\tilde{\mathbf{T}} \leftarrow [\,]$, $\mathbf{R} \leftarrow \mathbf{T}$; precompute pairwise distance matrix $D(\cdot,\cdot)$;
// Initialization stage: select the first representative token;
$k \leftarrow \arg\max_{i \in \mathbf{R}} \frac{1}{|\mathbf{R}|} \sum_{j \in \mathbf{R}} d(i,j)$;
Move k from $\mathbf{R}$ to $\tilde{\mathbf{T}}$;
// Iterative stage: select tokens that maximize diversity;
while $|\tilde{\mathbf{T}}| < \tilde{m}$ **do**
  for $i \in \mathbf{R}$ **do**
    $d_{\min}(i) \leftarrow \min_{j \in \tilde{\mathbf{T}}} d(i,j)$
  $k \leftarrow \arg\max_{i \in \mathbf{R}} d_{\min}(i)$;
  Move k from $\mathbf{R}$ to $\tilde{\mathbf{T}}$;
return $\tilde{\mathbf{T}}$

that maximizes token diversity, effectively suppressing redundancy and promoting representativeness.

However, the exact enumeration of all possible subsets is computationally intractable due to combinatorial explosion [3]. To address this, we adopt a greedy approximation [2, 56] that iteratively selects tokens to maximize diversity with negligible computational overhead and without additional training.

Greedy selection procedure. As summarized in Algorithm 1, the subset $\tilde{\mathbf{T}}$ is initialized as empty, and all tokens in $\mathbf{T}$ are placed in the candidate set $\mathbf{R}$. In the initialization step, the first token added to $\tilde{\mathbf{T}}$ is the one with the largest average distance to all other tokens in $\mathbf{R}$, ensuring that the process starts from a representative and globally distinct token. In subsequent iterations, each new token is chosen from $\mathbf{R}$ such that it maximizes the minimum distance to all tokens already in $\tilde{\mathbf{T}}$, encouraging maximal dissimilarity and hence diversity. This process repeats until $\tilde{\mathbf{T}}$ reaches the target size $\tilde{m}$. To accelerate computation, pairwise distances are precomputed once using matrix multiplication and reused during selection.

3.4 Pathology Token Scaling and Slide-Level Self-Supervised Pretraining

To enable scalable slide-level pretraining while retaining fine-grained spatial information, we employ our selective token retention strategy (**TokRet**, Section 3.3) to reduce redundancy among spatial tokens. Unlike conventional [CLS]-only representations, which compress a 14×14 tile into a single token ($196 \rightarrow 1$, $\sim 200\times$ compression), TokRet preserves approximately 10% of spatial tokens per region. This substantially increases token coverage while keeping computation tractable as shown in Figure 1.

As an illustrative example, a WSI region of size 7168×7168 pixels with a patch size of 16×16 pixels produces

$$N_{\text{tokens}} = \frac{7168 \times 7168}{16 \times 16} = 448 \times 448 = 200,704 \tag{8}$$

tokens. Retaining 10% of these tokens still results in about $\tilde{m} \approx 20{,}000$ tokens per region. When scaled to full-slide images, the token resolution readily surpasses 7,168, emphasizing the demand for a more computationally efficient architecture for over 20,000 tokens sequence modeling.

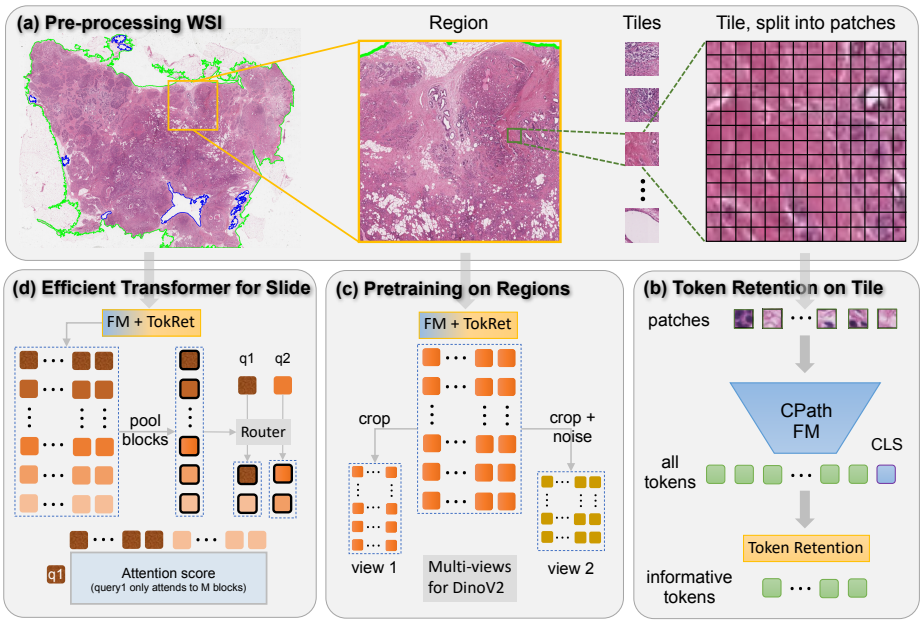


Fig. 2: Overall framework. (a) Pre-processing WSI Caption: (a) Pre-processing of whole slide images (WSI) step-by-step into regions, tiles and patches. (b) Token Retention on Tile: Token Retention (TokRet) mechanism applied to patches within a tile, using Computational Pathology (CPath) Foundation Model (FM) for informative token selection. (c) Pretraining on Regions: Pretraining is performed based on tokens from regions via FM+TokRet, two different views are generated for DinoV2 self-distillation. (d) Efficient Transformer for Slide: Efficient Transformer architecture for slide-level processing, utilizing FM + TokRet and attention routing to specific blocks. Each query token only need to attend to key tokens within M specific blocks, improving training and inference efficiency.

Efficient Transformer Encoding. To handle the large token set generated after TokRet (potentially up to ∼200k tokens per full-slide image), we adopt a long-context Transformer with near-linear attention complexity. Inspired by Mixture-of-Block-Attention [44] and Native-Sparse-Attention [79], we partition

the retained token set $\mathbf{T}_r$ into non-overlapping blocks $\{\mathbf{B}_1, \ldots, \mathbf{B}_K\}$. For each query token $\mathbf{q}_i$, we first compute a coarse attention score with each block:

$$\tilde{s}_{i,k} = \mathbf{q}_i^\top \operatorname{mean}(\mathbf{K}_k), \tag{9}$$

where $\mathbf{K}_k$ is the key matrix of block $\mathbf{B}_k$. Only the top-M blocks with highest $\tilde{s}_{i,k}$ are selected for fine-grained attention computation (as shown in Figure 2d, this process are denoted as Router):

$$\mathbf{z}_i = \sum_{k \in \mathrm{TopM}(i)} \operatorname{softmax}\big(\mathbf{q}_i^\top \mathbf{K}_k / \sqrt{d}\big)\mathbf{V}_k, \tag{10}$$

where $\mathbf{V}_k$ is the value matrix of block k. This approach reduces quadratic complexity while maintaining accurate modeling of relevant local and global dependencies.

For pretraining efficiency, we sample moderately sized regions (e.g., 7168 × 7168 pixels, producing ∼20k tokens after TokRet) to allow large batch sizes in pretraining process. To ensure generalization to larger downstream full-slide images, we integrate Rotary Positional Embedding (RoPE) [61], encoding relative positions. RoPE enables the retained token embeddings to adapt seamlessly to larger spatial extents while preserving spatial consistency.

This combination of block-based sparse attention and RoPE ensures that the Transformer can efficiently process 100k to 200k tokens per slide while preserving fine-grained morphological patterns critical for downstream WSI tasks.

Slide-Level Self-Supervised Pretraining. For slide-level pretraining, we employ a DINOv2 multi-view consistency self-distillation objective [50]. Each retained region generates two augmented views in token space (as shown Figure 2c), $\mathbf{T}_r^{(1)}$ and $\mathbf{T}_r^{(2)}$, which are processed by a student network $\mathcal{F}_\theta$ and a momentum teacher network $\mathcal{F}_{\theta'}$, respectively. The student network is trained to align its output distribution with that of the teacher. We omit details here, please check Appendix source code.

3.5 Overall Framework

Our framework integrates selective token retention, efficient Transformer encoding and pretraining, as shown in Figure 2. The pipeline is as follows:

1. **Token Retention.** Full-slide images are divided into regions. Feature tokens are extracted via a foundation model, and **TokRet** retains the top 10% of tokens per tile, preserving fine-grained information.
2. **Path-Token Scalable Pretraining.** Retained tokens (up to 20k per region) are fed into an efficient long-context Transformer with block-based sparse attention, enabling scalable modeling of large token sets. Positional information is encoded, and slide-level representations are learned via DINOv2 pretraining.
3. **Downstream Adaptation.** The pretrained slide-level model are adapted via further fine-tuning for downstream tasks such as diagnosis, prognosis and biomarker prediction, maintaining both local detail and global context for specific tasks.

4 Experiments

In this section, we evaluate the performance of the proposed method and compare it with various baselines. Additionally, we conduct ablation studies to further analyze its effectiveness.

Table 1: Slide-Level Classification based on FM. The results in the first-row are all trained on **UNI-1** CLS features, while the second-row we include some recent stronger FMs. The cyan rows are our methods. The **bold** and underline denote the best and second-best result, respectively.

	Tumor Grading		Biomarker Prediction		Gene Mutation	
	BRACS		BCNB ER		TCGA-LUAD TP53	
Method	F1	AUC	F1	AUC	F1	AUC
ABMIL [28]	0.692±0.03	0.875±0.02	0.754±0.08	0.857±0.07	0.646±0.04	0.701±0.07
CLAM-SB [46]	0.640±0.05	0.844±0.03	0.760±0.07	0.850±0.06	0.621±0.07	0.669±0.07
DTFD-MIL [80]	0.655±0.03	0.878±0.02	0.769±0.09	0.868±0.08	0.663±0.04	0.719±0.06
TransMIL [58]	0.592±0.03	0.859±0.02	0.690±0.06	0.810±0.08	0.608±0.04	0.674±0.08
MambaMIL [76]	0.650±0.04	0.854±0.03	0.770±0.06	0.860±0.07	0.631±0.08	0.679±0.06
GMMamba [86]	0.670±0.04	0.893±0.03	0.784±0.08	0.873±0.07	0.678±0.04	0.734±0.05
PathVQ [36]	0.730±0.02	0.902±0.01	0.782±0.07	0.872±0.05	0.684±0.03	0.733±0.06
GigaPath [75]	0.677±0.03	0.862±0.03	0.791±0.07	0.875±0.06	0.670±0.04	0.738±0.06
TITAN [19]	0.696±0.04	0.891±0.01	**0.813±0.06**	**0.905±0.05**	0.702±0.03	0.765±0.06
UNI-2 (ABMIL)	0.698±0.03	0.887±0.02	0.770±0.07	0.861±0.06	0.661±0.03	0.729±0.06
Ours +UNI-1	**0.779±0.03**	**0.914±0.03**	0.801±0.05	0.890±0.06	0.714±0.03	0.763±0.04
Ours +Conch-1.5	0.767±0.05	0.910±0.05	0.809±0.07	0.898±0.05	**0.725±0.05**	**0.771±0.03**

4.1 Pretraining

Data Description We perform large-scale self-supervised pretraining using all diagnostic WSIs from the HistAI [49] cohorts together with our private WSIs (about 10k). Each WSI is partitioned into non-overlapping regions of 7168×7168 pixels. After quality control, a total of 98,318 regions are retained for pretraining. Patch embeddings are then extracted from smaller 448×448 tiles using a frozen pathology foundation model (UNI-1 [11], we also done Conch-1.5 [18] for comparison).

Implementation Details To reduce redundancy while preserving fine-grained spatial cues, the proposed **TokRet** module retains approximately 10% of the most informative tokens per region. This yields about 20,480 tokens per region, resulting in roughly **2 billion tokens** in total for slide-level pretraining.

The retained tokens are processed by a lightweight 6-layer Transformer with **block-based sparse attention** [44]. Specifically, attention is computed in a

coarse-to-fine manner: each query first interacts with coarse-grained key blocks to estimate inter-block similarity, and detailed token-level attention is computed only within the top-M most relevant blocks (with block size set to 512 and $M = 4$). This strategy reduces attention complexity to near-linear while maintaining long-range context modeling. Positional encoding is handled by a 2-D implementation of RoPE [61], allowing the model to generalize from 7168×7168 regions to higher-resolution WSIs during downstream adaptation. For all retained tokens originating from the same tile, identical position indices are assigned, with intra-patch variation captured implicitly by the tile-level ViT features.

Self-supervised pretraining follows the DINOv2 framework. We train for 50 epochs on 8 NVIDIA A100 GPUs with a batch size of 16 per GPU, using the AdamW optimizer (base learning rate 5×10^{-4}, cosine decay schedule, and weight decay 4×10^{-2}). Data augmentations, including random cropping, horizontal flipping and gaussian noise in feature space, are applied at the region level to generate diverse views.

For more details of pre-training including data, codes, model structure, hyper-parameters and training logs, please check Appendix.

4.2 Downstream Tasks

We primarily focus on WSI classification and survival prediction.

WSI Classification Tasks We evaluate our method on three kinds of WSI classification tasks including: **(1) Tumor Subtyping**. BReAst Carcinoma Subtyping (BRACS) [5] collect H&E stained Histology Images, containing 547 WSIs for three lesion types, i.e., benign, malignant and atypical, which are further subtyped into seven categories. Here, since the WSIs number is limited, we only perform three class subtyping. **(2) Biomarker prediction**. BCNB [74] (n = 1038), a dataset of early breast cancer core-needle biopsy WSI is used to predict tumor clinical characteristic estrogen receptor (ER). **(3) Gene mutation prediction**. A subset of The Cancer Genome Atlas (TCGA) [74] is used (TCGA-LUAD, n = 469) to predict TP53 gene mutation from lung adenocarcinoma WSIs.

The data processing and embedding procedure are identical to the region-based approach but are applied at the WSI level. The (x, y) coordinates of tiles are also stored to facilitate positional encoding in the WSI-Transformer. During fine-tuning, the overall WSI model are trained with a batch size of 1 for 20 epochs. The learning rate is fixed at 1×10^{-4}, with a weight decay of 1×10^{-4}, using the AdamW optimizer with default settings.

For more evaluation dataset and protocol details, please refer to Appendix.
Compared Baselines: Since our method primarily focuses better pathology FM and efficient Transformer, we compare it against various WSI analysis models with different architectural designs:

(1) Simple MILs architectures: ABMIL [28], DSMIL [35] (introduces a max-pooling branch alongside the attention mechanism), and DTFD-MIL [80] (employs

sub-bags for hierarchical learning). All these results are based on UNI-1 tile-level CLS features.

(2) Efficient Transformer/attention based MILs: TransMIL [58] (leveraging Nyström self-attention [73] for computational efficiency) and Mamba based MIL method [76, 86]. All these results are based on UNI-1 tile-level CLS features.

(3) Pathology FMs: like GigaPath (a 12-layers WSI-Transformer (efficiently implemented using LongViT [68]), pretrained on large-scale private data via MAE [25], with a [CLS] token as tile feature), TITAN [18, 75] (a 6-layers WSI-Transformer with 2D-ALiBi positional encoding [37, 51], pretrained on large-scale private data using iBOT [87], with Conch-v1.5 as the tile feature extractor ([CLS] token).), and UNI-2 are also included.

(4) Token feature dimension compression method, PathVQ [36] is included because of similar spatial token information preservation purpose.

Works pursuing orthogonal goals (e.g., overfitting mitigation, hard-instance mining) are not part of our primary comparison [17, 42, 53, 62, 64, 76, 83, 84, 88].

For all the experiments, we report the macro-AUC and macro-F1 scores (over five-runs or five-fold cross validation) because of class imbalance.

WSI Classification Results Analysis. Table 1 summarizes the slide-level classification performance across tumor grading (BRACS), biomarker prediction (BCNB ER), and gene mutation prediction (TCGA-LUAD TP53). Across all three tasks, traditional MIL baselines such as ABMIL, CLAM-SB, MambaMIL, and TransMIL exhibit moderate performance, while recent foundation-model–based pipelines (e.g., GigaPath and TITAN) provide stronger results.

Our proposed **PathTokScale** achieves the best or second-best performance in *every* metric, delivering consistent and substantial improvements over all competing approaches. Notably, it surpasses strong baselines built on modern FMs (e.g., GigaPath, TITAN, and UNI-2–based ABMIL) without increasing model size or requiring massive additional tile-scale pretraining. Our method also show consistent improvement on different tile-level backbones (on both UNI-1 and Conch-1.5).

These improvements highlight an important insight: the bottleneck of existing pathology FMs largely stems from ineffective token retention and the resulting information loss in the global representation. While prior FMs attempt to scale performance primarily by expanding data size (100k–1M slides) or model capacity (2–6× parameters in both tile [89] or slide level [75]), their slide-level performance still saturates. In contrast, **PathTokScale** directly mitigates the representation bottleneck by preserving a more complete set of semantically critical tokens, enabling markedly better utilization of existing FM features.

Overall, these results demonstrate that rethinking token preservation—rather than simply scaling data or model size—is key to unlocking the next stage of scalability for pathology FMs.

WSI Survival Prediction For paper length, please check dataset details and comparison baselines description in Appendix. The model is trained using the

Table 2: Survival prediction Results based FM measuring patient disease-specific survival. All methods in Prototype and MIL use **UNI-1** CLS features [11]. Best performance in **bold**, second best underlined.

	TCGA	BRCA	CRC	BLCA	UCEC	KIRC	Average
Prototype	H2T [67]	0.672±0.07	0.639±0.11	0.566±0.05	0.715±0.09	0.703±0.11	0.659
	OT [48]	0.755±0.06	0.622±0.09	0.603±0.04	0.747±0.08	0.695±0.09	0.684
	PANTHER [60]	**0.758±0.06**	0.665±0.10	0.612±0.07	0.757±0.10	0.716±0.10	0.702
MIL	AttnMISL [77]	0.627±0.08	0.639±0.10	0.485±0.06	0.581±0.12	0.649±0.09	0.596
	ILRA [72]	0.649±0.10	0.555±0.10	0.550±0.04	0.632±0.02	0.637±0.14	0.605
	TransMIL [58]	0.612±0.07	0.684±0.06	0.595±0.06	0.695±0.08	0.671±0.10	0.651
	ABMIL [28]	0.633±0.06	0.612±0.08	0.540±0.07	0.671±0.08	0.691±0.08	0.629
	PathVQ [36]	0.655±0.05	0.649±0.12	0.608±0.05	0.721±0.10	0.760±0.08	0.679
FMs	CHIEF [70]	0.737±0.04	0.680±0.08	0.599±0.02	0.758±0.10	0.736±0.06	0.702
	GigaPath [75]	0.687±0.08	0.628±0.08	0.589±0.05	0.779±0.10	0.751±0.07	0.687
	TITAN [19]	0.713±0.04	0.710±0.11	**0.657±0.05**	**0.789±0.09**	0.774±0.06	**0.729**
	UNI-2 (ABMIL)	0.614±0.02	0.618±0.11	0.539±0.08	0.672±0.08	0.659±0.11	0.620
PathTokScale	Ours +UNI-1	0.712±0.07	0.699±0.09	0.635±0.04	0.769±0.06	0.767±0.09	0.716
	Ours +Conch-1.5	0.717±0.04	**0.719±0.07**	0.655±0.05	0.784±0.08	**0.778±0.07**	0.731

negative log-likelihood (NLL) loss and evaluated using the c-index with 5-fold cross validation (the result of last epoch is reported).

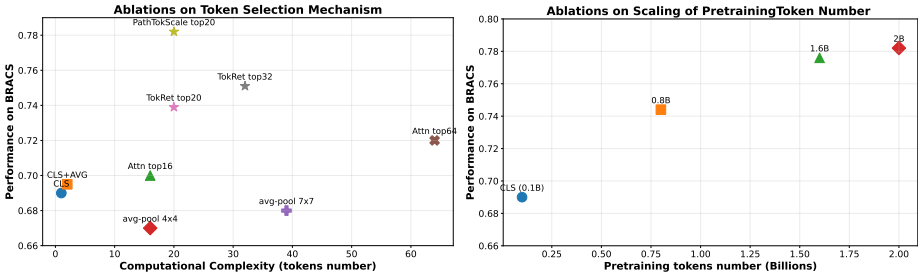
WSI Survival Prediction Results Analysis. Table 2 summarizes the performance across survival cohorts. Prototype-based and MIL-based approaches (e.g., H2T, ILRA, TransMIL) provide moderate C-index values but remain limited by coarse pooling and substantial information collapse, which is particularly detrimental for survival modeling where subtle morphological cues are strongly tied to patient risk. Although PathVQ improves over MIL baselines by retaining more informative instances, the gains remain modest, indicating that token dimension reduction struggles to preserve prognostically critical features.

FM-based pipelines (CHIEF, GigaPath—pretrained on over 170k slides, TITAN—pretrained on over 330k slides, and UNI-2+ABMIL) achieve strong overall performance, but their gains plateau despite large-scale pretraining or increased model capacity. This trend supports our hypothesis that the [CLS]-centric representation bottleneck fundamentally limits current FMs from capturing heterogeneous risk signals. In contrast, our **PathTokScale**, trained on only 30k slides (about 10% of TITAN's data), achieves comparable performance without additional pretraining or larger backbones. These results suggest that improving token retention—rather than merely scaling data or parameters—is crucial for reliable and generalizable survival prediction from WSIs.

4.3 Ablation Study

4.4 Ablation on Token Selection Mechanism

We first investigate how the number of retained tokens influences both computational cost and downstream performance. As shown in Figure 3a, reducing K from 32 to 20 significantly while maintaining strong accuracy, demonstrating that a small subset of informative spatial tokens already captures most of the



(a) Comparison of different token selection mechanisms on BRACS, showing performance versus computational complexity. TokRet and PathTokScale methods achieve higher accuracy with fewer retained tokens.

(b) The scaling analysis on pretraining tokens number. When scaling up tokens number via TokRet, the performance are gradually improved. Noted that the CLS-only approach can be seen as a starting point for scaling.

Fig. 3: Ablation study on token selection mechanism and scaling analysis on pretraining token.

discriminative signal. We provide more K selection in Appendix with findings that performance degrades only when K becomes extremely small (e.g., K $\leq$ 4), confirming that TokRet successfully identifies representative tokens that preserve the global feature distribution.

In Figure 3a, we first show vanilla baselines CLS-only and CLS+AVG (concatenate CLS and average-pooling single vector). We further compare our greedy dissimilarity-based selection with two commonly-used baselines: (1) attention Top-K, where tokens are ranked by self-attention magnitude, and (2) average pooling, which compresses all spatial tokens into a single vector. While attention Top-K moderately improves over pure [CLS] modeling, it remains less effective and less stable than TokRet due to its reliance on attention heuristics. Average pooling consistently underperforms, validating that token diversity—not token averaging—is essential for capturing fine-grained morphological patterns. TokRet achieves the best balance between efficiency and representational richness across all tested K.

Sensitivity on Top-K Number We show the sensitivity on Top-K Number of token retention in Figure 4. when $K \leq 4$, the performance are dropped compared to [CLS]. But when more tokens ($K \geq 8$) are preserved, the performance are steadily improved.

For **more ablation experiments** including sensitivity on Top-K Number of token retention, sparse-attention block size and top-M, scaling analysis on different token number and model size, memory and computational cost analysis of our method, please check Appendix. Especially, for sparse-attention block size and top-M, increasing block size from 512 and Top-M from 4 to larger number does not improve performance but introduce too much computation cost.

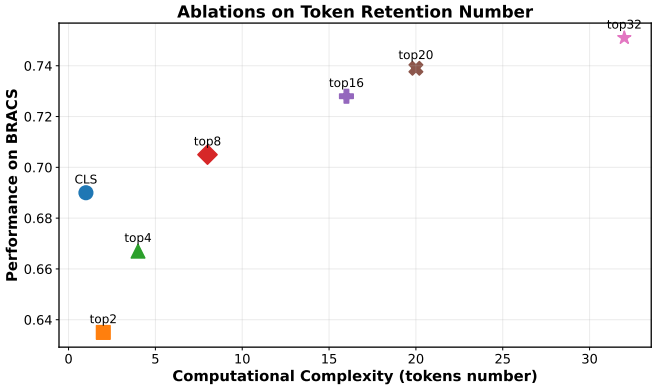


Fig. 4: Comparison of different token retention top$K$ number on BRACS, showing performance versus computational complexity. When $K \leq 4$, the performance are dropped compared to [CLS].

5 Conclusion

In this work, we revisited the representational bottlenecks that arise from the prevalent reliance on the [CLS] token in pathology foundation models. We demonstrated that spatial tokens contain rich, fine-grained cues essential for modeling subtle morphological variations in whole slide images, yet their direct use is computationally prohibitive due to the massive token counts inherent to gigapixel pathology data. To overcome this challenge, we introduced TokRet, a plug-and-play selective token retention and compression module that preserves only the most informative spatial tokens while removing redundancy. Built upon TokRet, we further proposed PathTokScale, an efficient slide-level pretraining framework that scales to billion-token training regimes—substantially surpassing the capacity of existing computational pathology approaches.

Overall, TokRet and PathTokScale address a critical scalability barrier in computational pathology and open a promising direction toward efficient, high-resolution, and high-capacity foundation models for whole slide image analysis.

Acknowledgements

This study was partially supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant 2025SDXHDX0003), the National Natural Science Foundation of China (Grant No.62506306), and National Key R&D Program of China (Grant No. 2025YFC3508600). Furthermore, essential technical support was provided by the D-PathAI platform, including its hardware and software, which was developed by Hangzhou Dipath Technology Co., Ltd.

References

1. Albastaki, S., Sohail, A., Ganapathi, I.I., Alawode, B., Khan, A., Javed, S., Werghi, N., Bennamoun, M., Mahmood, A.: Multi-resolution pathology-language pre-training model with text-guided visual representation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25907–25919 (2025)
2. Alvar, S.R., Singh, G., Akbari, M., Zhang, Y.: Divprune: Diversity-based visual token pruning for large multimodal models. arXiv preprint arXiv:2503.02175 (2025)
3. Blum, C., Puchinger, J., Raidl, G.R., Roli, A.: Hybrid metaheuristics in combinatorial optimization: A survey. Applied soft computing **11**(6), 4135–4151 (2011)
4. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022)
5. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., Pietro, G.D., Bonito, M.D., Foncubierta, A., Botti, G., Gabrani, M., Feroce, F., Frucci, M.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images (2021)
6. Cai, J., Zhu, C., Cui, C., Li, H., Wu, T., Zhang, S., Yang, L.: Generalizing nucleus recognition model in multi-source ki67 immunohistochemistry stained images via domain-specific pruning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. pp. 277–287. Springer (2021)
7. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. CoRR **abs/2104.14294** (2021), `https://arxiv.org/abs/2104.14294`
9. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: European Conference on Computer Vision. pp. 19–35. Springer (2024)
10. Chen, R.J., al, e.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: CVPR. pp. 16144–16155 (June 2022)
11. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
12. Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F.K., Mahmood, F.: Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks (2021)
13. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4025 (October 2021)
14. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
15. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. arXiv preprint arXiv:2104.02057 (2021)

16. Chen, X., Qiu, P., Zhu, W., Wang, H., Li, H., Dong, X., Sun, X., Yu, X., Wang, Y., Razi, A., et al.: Cracking instance jigsaw puzzles: An alternative to multiple instance learning for whole slide image analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21353–21363 (2025)
17. CUI, Y., Liu, Z., Liu, X., Liu, X., Wang, C., Kuo, T.W., Xue, C.J., Chan, A.B.: Bayes-MIL: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=_geIwiOyUhZ
18. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., Williamson, D.F.K., Chen, B., Almagro-Perez, C., Doucet, P., Sahai, S., Chen, C., Komura, D., Kawabe, A., Ishikawa, S., Gerber, G., Peng, T., Le, L.P., Mahmood, F.: Multimodal whole slide foundation model for pathology (2024), https://arxiv.org/abs/2411.19666
19. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., et al.: Multimodal whole slide foundation model for pathology. arXiv preprint arXiv:2411.19666 (2024)
20. Dong, J., Jiang, J., Jiang, K., Li, J., Zhang, Y.: Fast and accurate gigapixel pathological image classification with hierarchical distillation multi-instance learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30818–30828 (2025)
21. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Kain, A.M., Saillard, C., Schiratti, J.B.: Scaling self-supervised learning for histopathology with masked image modeling. medRxiv (2023). https://doi.org/10.1101/2023.07.21.23292757, https://www.medrxiv.org/content/early/2023/07/26/2023.07.21.23292757
22. Ganguly, A., Chatterjee, D., Huang, W., Zhang, J., Yurovsky, A., Johnson, T.S., Chen, C.: Merge: Multi-faceted hierarchical graph-based gnn for gene expression prediction from whole slide histopathology images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15611–15620 (2025)
23. Ghezloo, F., Seyfioglu, M.S., Soraki, R., Ikezogwo, W.O., Li, B., Vivekanandan, T., Elmore, J.G., Krishna, R., Shapiro, L.: Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. arXiv preprint arXiv:2502.08916 (2025)
24. Guo, Z., Xiong, C., Ma, J., Sun, Q., Feng, L., Wang, J., Chen, H.: Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15590–15600 (2025)
25. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. CoRR **abs/2111.06377** (2021), https://arxiv.org/abs/2111.06377
26. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022)
27. Hou, X., Jiang, C., Kondepudi, A., Lyu, Y., Chowdury, A., Lee, H., Hollon, T.C.: A self-supervised framework for learning whole slide representations. arXiv preprint arXiv:2402.06188 (2024)
28. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/ilse18a.html

29. Jaume, G., Oldenburg, L., Vaidya, A., Chen, R.J., Williamson, D.F.K., Peeters, T., Song, A.H., Mahmood, F.: Transcriptomics-guided slide representation learning in computational pathology (2024), https://arxiv.org/abs/2405.11618
30. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020), https://arxiv.org/pdf/2001.08361.pdf
31. Kapse, S., Pati, P., Yellapragada, S., Das, S., Gupta, R.R., Saltz, J., Samaras, D., Prasanna, P.: Gecko: Gigapixel vision-concept contrastive pretraining in histopathology. arXiv preprint arXiv:2504.01009 (2025)
32. Kuo, C.C., Glover, F., Dhir, K.S.: Analyzing and modeling the maximum diversity problem by zero-one programming. Decision Sciences **24**(6), 1171–1185 (1993)
33. Lazard, T., Lerousseau, M., Decencière, E., Walter, T.: Giga-ssl: Self-supervised learning for gigapixel images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4305–4314 (2023)
34. Lazard, T., Lerousseau, M., Decencière, E., Walter, T.: Giga-ssl: Self-supervised learning for gigapixel images (2022), https://arxiv.org/abs/2212.03273
35. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2021)
36. Li, H., Shui, Z., Zhang, Y., Zhu, C., Yang, L.: Pathvq: Reforming computational pathology foundation model for whole slide image analysis via vector quantization. arXiv preprint arXiv:2503.06482 (2025)
37. Li, H., Zhang, Y., Chen, P., Shui, Z., Zhu, C., Yang, L.: Rethinking transformer for long contextual histopathology whole slide image analysis. arXiv preprint arXiv:2410.14195 (2024)
38. Li, H., Zhu, C., Zhang, Y., Sun, Y., Shui, Z., Kuang, W., Zheng, S., Yang, L.: Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7454–7463 (June 2023)
39. Li, W., Yuan, Y., Liu, J., Tang, D., Wang, S., Qin, J., Zhu, J., Zhang, L.: Tokenpacker: Efficient visual projector for multimodal llm. International Journal of Computer Vision pp. 1–19 (2025)
40. Li, X., Cui, Y., Li, J., Chan, A.B.: Advancing multiple instance learning with continual learning for whole slide imaging. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20800–20809 (2025)
41. Liang, Y., Lyu, X., Chen, W., Ding, M., Zhang, J., He, X., Wu, S., Xing, X., Yang, S., Wang, X., et al.: Wsi-llava: A multimodal large language model for whole slide image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22718–22727 (2025)
42. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19830–19839 (2023)
43. Lin, Z., Lin, M., Lin, L., Ji, R.: Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5334–5342 (2025)
44. Lu, E., Jiang, Z., Liu, J., Du, Y., Jiang, T., Hong, C., Liu, S., He, W., Yuan, E., Wang, Y., et al.: Moba: Mixture of block attention for long-context llms. arXiv preprint arXiv:2502.13189 (2025)

45. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)
46. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
47. Ma, Y., An, B., Shen, A., Yuan, M., Duan, M., Wang, M.: Flow-mil: Constructing highly-expressive latent feature space for whole slide image classification using normalizing flow. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23561–23570 (2025)
48. Mialon, G., Chen, D., d'Aspremont, A., Mairal, J.: A trainable optimal transport embedding for feature aggregation and its relationship to attention. arXiv preprint arXiv:2006.12065 (2020)
49. Nechaev, D., Pchelnikov, A., Ivanova, E.: Histai: An open-source, large-scale whole slide image dataset for computational pathology (2025), https://arxiv.org/abs/2505.12120
50. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
51. Press, O., Smith, N.A., Lewis, M.: Train short, test long: Attention with linear biases enables input length extrapolation. arXiv preprint arXiv:2108.12409 (2021)
52. Qiu, Z., Chao, H., Lin, T., Chang, W., Yang, Z., Jiao, W., Shen, Y., Zhang, Y., Yang, Y., Liu, W., et al.: Bridging local inductive bias and long-range dependencies with pixel-mamba for end-to-end whole slide image analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22738–22747 (2025)
53. Qu, L., xiaoyuan Luo, Wang, M., Song, Z.: Bi-directional weakly supervised knowledge distillation for whole slide image classification. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=JoZyVgp1hm
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
55. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
56. Resende, M.G., Martí, R., Gallego, M., Duarte, A.: Grasp and path relinking for the max–min diversity problem. Computers & Operations Research **37**(3), 498–508 (2010)
57. Shang, Y., Cai, M., Xu, B., Lee, Y.J., Yan, Y.: Llava-prumerge: Adaptive token reduction for efficient large multimodal models. arXiv preprint arXiv:2403.15388 (2024)
58. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., zhang, y.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 2136–2147. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper/2021/file/10c272d06794d3e5785d5e7c5356e9ff-Paper.pdf
59. Shou, Y., Cao, X., Yan, P., Hui, Q., Zhao, Q., Meng, D.: Graph domain adaptation with dual-branch encoder and two-level alignment for whole slide image-based survival prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19925–19935 (2025)

60. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11566–11578 (2024)
61. Su, J., Lu, Y., Pan, S., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding (2021)
62. Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple instance learning framework with masked hard instance mining for whole slide image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4078–4087 (2023)
63. Tang, W., Qin, R., Fang, H., Zhou, F., Chen, H., Li, X., Cheng, M.M.: Revisiting end-to-end learning with slide-level supervision in computational pathology. arXiv preprint arXiv:2506.02408 (2025)
64. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11343–11352 (2024)
65. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: Review the cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia **2015**(1), 68–77 (2015)
66. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
67. Vu, Q.D., Rajpoot, K., Raza, S.E.A., Rajpoot, N.: Handcrafted histological transformer (h2t): Unsupervised representation of whole slide images. Medical image analysis **85**, 102743 (2023)
68. Wang, W., Ma, S., Xu, H., Usuyama, N., Ding, J., Poon, H., Wei, F.: When an image is worth 1,024 x 1,024 words: A case study in computational pathology. arXiv preprint arXiv:2312.03558 (2023)
69. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W., Han, X.: Transpath: Transformer-based self-supervised learning for histopathological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 186–195. Springer (2021)
70. Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature **634**(8035), 970–978 (2024)
71. Wu, J., Chen, M., Ke, X., Xun, T., Jiang, X., Zhou, H., Shao, L., Kong, Y.: Learning heterogeneous tissues with mixture of experts for gigapixel whole slide images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5144–5153 (2025)
72. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=01KmhBsEPFO
73. Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V.: Nyströmformer: A nyström-based algorithm for approximating self-attention. Proceedings of the AAAI Conference on Artificial Intelligence (2021)
74. Xu, F., Zhu, C., Tang, W., Wang, Y., Zhang, Y., Li, J., Jiang, H., Shi, Z., Liu, J., Jin, M.: Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. Frontiers in oncology **11**, 759007 (2021)

75. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature pp. 1–8 (2024)
76. Yang, S., Wang, Y., Chen, H.: Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 296–306. Springer (2024)
77. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis **65**, 101789 (2020)
78. Ye, W., Wu, Q., Lin, W., Zhou, Y.: Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 22128–22136 (2025)
79. Yuan, J., Gao, H., Dai, D., Luo, J., Zhao, L., Zhang, Z., Xie, Z., Wei, Y., Wang, L., Xiao, Z., et al.: Native sparse attention: Hardware-aligned and natively trainable sparse attention. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 23078–23097 (2025)
80. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. ArXiv **abs/2203.12081** (2022)
81. Zhang, J., Kapse, S., Ma, K., Prasanna, P., Saltz, J., Vakalopoulou, M., Samaras, D.: Prompt-mil: Boosting multi-instance learning schemes via task-specific prompt tuning (2023)
82. Zhang, J., Nguyen, A.T., Han, X., Trinh, V.Q.H., Qin, H., Samaras, D., Hosseini, M.S.: 2dmamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3583–3592 (2025)
83. Zhang, Y., Li, H., Sun, Y., Zheng, S., Zhu, C., Yang, L.: Attention-challenging multiple instance learning for whole slide image classification. arXiv preprint arXiv:2311.07125 (2023)
84. Zhang, Y., Shui, Z., Sun, Y., Li, H., Li, J., Zhu, C., Zheng, S., Yang, L.: Adr: Attention diversification regularization for mitigating overfitting in multiple instance learning based whole slide image classification. arXiv preprint arXiv:2406.15303 (2024)
85. Zheng, T., Jiang, K., Xiao, Y., Zhao, S., Yao, H.: M3amba: Memory mamba is all you need for whole slide image classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15601–15610 (2025)
86. Zheng, T., Yao, H., Jiang, K., Xiao, Y., Zhao, S.: Gmmamba: Group masking mamba for whole slide image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9935–9944 (2025)
87. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)
88. Zhu, W., Chen, X., Qiu, P., Sotiras, A., Razi, A., Wang, Y.: Dgr-mil: Exploring diverse global representation in multiple instance learning for whole slide image classification. In: European Conference on Computer Vision. pp. 333–351. Springer (2024)
89. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)