Scalable Whole-Slide Vision-Language Modeling with Learned Token Pruning

Ali Kerem Bozkurt[†,⋆], **Baris Cem Bakay**[†,⋆], **Ibrahim Kulac**[*,⋆], **Cigdem Gunduz Demir**[†,⋆], **Erkut Erdem**[‡,⋆], **Aykut Erdem**[†,⋆]

[†] Department of Computer Engineering, Koç University, Istanbul, Turkey
[*] Department of Pathology, School of Medicine, Koç University, Istanbul, Turkey
[‡] Department of Computer Engineering, Hacettepe University, Ankara, Turkey
[⋆] KUIS AI Center, Koç University, Istanbul, Turkey

Abstract

Efficient modeling of whole-slide images (WSIs) is a central challenge in digital pathology. A single slide can expand into tens of thousands of patch tokens, pushing beyond the limits of standard transformer architectures and creating prohibitive computational costs. Existing foundational models employ efficient attention mechanisms, yet massive token counts remain a bottleneck. We propose SLIM (Slide-Level Interpretable Modeling with Token Pruning), a whole-slide vision–language framework that makes efficiency a core design principle by integrating token pruning into the slide representation stage. Starting from pretrained CONCH v1.5 patch embeddings, a LongNet-based encoder models ultra-long sequences while Cropr modules progressively discard low-utility tokens. Unlike token compression or merging, pruning directly shortens sequences, lowering memory and latency while preserving diagnostically relevant context. The pruning signal also offers interpretability, echoing how pathologists scan slides by ignoring background and focusing on salient tissue. For multimodal alignment, we adopt a CLIP-style contrastive objective with PubMedBERT as the text encoder, producing a compact joint space for retrieval and classification. Experiments on TCGA and EBRAINS show that pruning achieves a favorable efficiency–accuracy trade-off: our model matches or exceeds the performance of scale-heavy baselines such as Prov-GigaPath, while operating at an order of magnitude lower cost. Our results establish token pruning as a practical and interpretable strategy for scalable whole-slide modeling.

1 Introduction

Whole-slide images (WSIs) are central to digital pathology, providing detailed tissue morphology at gigapixel resolution and enabling tasks such as cancer grading, biomarker discovery, and prognosis. Their sheer scale, however, creates a fundamental computational barrier: a single WSI may contain millions of image patches, which translate into token sequences far beyond the capacity of standard transformer architectures. Because the cost of self-attention grows quadratically with sequence length, naïve approaches to slide-level modeling quickly become infeasible. This computational bottleneck is not only a technical challenge but also a practical obstacle for deploying foundation models in clinical and research settings, where efficient and reproducible methods are essential.

Recent work has sought to address these challenges by developing pathology-specific foundation models with long-context encoders and multimodal training objectives. Such models have demonstrated strong performance across diverse downstream tasks, including cancer subtyping, mutation prediction, and slide–report retrieval [20, 18, 7, 19]. In parallel, general-purpose multimodal language

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

models (MLLMs) such as CoCa [21] and LLaVA [12] have been adapted to the pathology domain, supporting interactive reasoning and visual question answering over WSIs [11, 3]. Together, these advances point to the promise of multimodal approaches for bridging image-level and language-level understanding in pathology. Yet, they also highlight two persistent limitations: (i) reliance on massive pretraining corpora and compute budgets, which hinders accessibility and reproducibility; and (ii) insufficient attention to efficiency, trust, and compliance concerns, which are especially critical in healthcare applications.

In this work, we revisit the challenge of WSI modeling from the perspective of token reduction and efficient sequence modeling—a line of research that has recently gained momentum in the broader transformer community. Our motivation is to explore whether these advances can be leveraged to make multimodal slide encoders more scalable, accessible, and ultimately more suitable for clinical integration. To this end, we introduce SLIM (Slide-Level Interpretable Modeling with Token Pruning), an efficient framework for whole-slide vision–language modeling.

SLIM builds on high-quality CONCH v1.5 [13] patch embeddings, which eliminate the need for patch-level pretraining. These embeddings are fed into a LongNet [6]-based slide encoder with two-dimensional positional encodings, allowing the model to capture global spatial structure across ultra-long sequences. To further improve scalability, we integrate Token Cropr [1], a token pruning mechanism that learns to identify and discard low-utility tokens. Unlike token compression or merging, pruning directly reduces sequence length while preserving task-relevant tokens, thereby lowering memory and latency costs without substantial architectural modifications.

Crucially, pruning also mirrors how pathologists interpret slides: they first scan tissue at low magnification, disregard visually uninformative regions, and concentrate on areas most relevant for diagnosis. In a similar way, SLIM allocates representational capacity to the most informative patches. This not only improves efficiency but also enhances interpretability, since the separation between discarded and preserved tokens reveals which regions the model considers salient for multimodal alignment, offering greater transparency into its decision-making process.

For multimodal alignment, we adopt a CLIP-style contrastive objective [15, 4] to align slide and report embeddings, using PubMedBERT [8] as a domain-specialized text encoder. By freezing the text encoder and training lightweight projection layers, we obtain a compact joint embedding space that supports retrieval and classification without requiring large-scale finetuning of language models. Our contribution is complementary to scale-heavy foundation models such as Prov-GigaPath [20], which pretrains tile and slide encoders on billions of tiles, and recent VQA-oriented models like TCP-LLaVA [14], which compress tokens before feeding them into a large language model. In contrast, we integrate token pruning into the slide representation stage itself, focusing on efficiency as a first-class design principle. This provides a practical path toward accessible and reproducible whole-slide vision-language models, making them feasible under realistic compute and data constraints.

In summary, our contribution is twofold. First, we introduce SLIM, a whole-slide vision–language modeling framework that emphasizes efficiency by addressing the core bottleneck of sequence length. Here, we extend the recently proposed Cropr [1] token pruning mechanism to digital pathology, showing how pruning can reduce computational cost while also offering interpretability by highlighting salient regions of the slide. Third, we demonstrate how this pruning-enabled slide encoder can be effectively aligned with clinical text using a CLIP-style contrastive objective with PubMedBERT, providing a practical recipe for scalable multimodal representation learning in digital pathology. Our experimental results demonstrate that this design achieves a favorable trade-off between efficiency and representation quality, highlighting the practicality of token pruning for WSI vision-language models.

2 Related Work

Whole-slide foundation and multimodal models. Digital pathology increasingly turns to *foundation models* to derive transferable slide-level representations across diseases, stains, scanners, and institutions. This need is driven by two facts: WSIs are gigapixel-scale, making end-to-end transformers impractical, and clinical tasks are diverse, so training one model per task is neither data- nor compute-efficient. The prevailing recipe extracts patch embeddings with a tile encoder and aggregates them with a slide encoder; recent work mainly differs in pretraining scale and multimodal supervision from pathology reports.

Prov-GigaPath [20] couples billion-tile pretraining (DINOv2) with a LongNet-style slide encoder and contrastive slide–report alignment, yielding strong generalization in subtyping, mutation prediction, and zero-shot diagnostics. PRISM [18] extends the multimodal direction by pairing ViT-based tile features (Virchow [19]) with hundreds of thousands of reports under contrastive and generative objectives, supporting both retrieval and report generation. TITAN [7] leverages CONCH v1.5 [13] features with a ViT slide encoder trained in multiple stages—self-supervision, synthetic captioning, and report alignment to enable long-context modeling. HistGen [9] proposes a hierarchical local-to-global encoder with a cross-modal context module for report generation and releases pre-extracted DINOv2 features with $\sim$7.7k WSI–report pairs plus a $>$55k-WSI pretraining corpus. These efforts establish the value of large-scale multimodal supervision in pathology, but they also share a limitation: progress has been driven primarily by *scale*, massive corpora and compute budgets, rather than by making slide representation intrinsically more efficient and easy to deploy.

Multimodal LLMs for digital pathology. The recent success of multimodal large language models (LLMs) has inspired efforts to bring language-driven reasoning to digital pathology. WSI-LLaVA [11] adapts the LLaVA framework to whole-slide images, introducing a multi-stage training pipeline for pathology-specific visual question answering (VQA) along with a dedicated benchmark. SlideChat [3] extends this direction by aligning a slide encoder with an LLM through instruction tuning and reporting strong performance on benchmark downstream tasks. These models illustrate the potential of LLM-based approaches to provide richer explanations and language-grounded reasoning in pathology. However, their emphasis lies in enabling dialogue and VQA capabilities rather than in learning efficient and transferable slide-level representations.

Efficient long-context Transformers and token reduction. Beyond pathology, efficiency in long-sequence transformers has advanced along two lines. (1) *Extended attention* (e.g., LongNet[6]) enlarges receptive fields via dilated attention for ultra-long contexts. (2) *Token reduction* shortens sequences via merging, selection, or pruning: ToMe [2] fuses similar tokens with minimal accuracy loss; TokenLearner [17] learns a compact, task-relevant tokens; and Token Cropr [1] prunes low-utility tokens, achieving 1.5–4$\times$ speedups with minimal degradation. Hybrid schemes such as PACT [5] combine pruning and merging to further reduce latency and memory in vision–language models. Yet, token-reduction strategies remain largely unexplored in WSI modeling, where efficiency is often pursued indirectly through larger pretraining. This gap highlights a mismatch between what long-context Transformers *can* process and what real-world clinical pipelines can *afford*.

The works reviewed above highlight three complementary directions: large-scale pathology foundation models that prioritize generalization through massive data and compute, multimodal LLMs that focus on reasoning and instruction following, and efficiency-oriented token reduction methods that have yet to be explored in digital pathology. Our contribution brings these strands together by introducing *learned token pruning* into the slide-representation stage. Starting from high-quality patch embeddings to avoid tile-level pretraining, we use a long-context slide encoder that dynamically prunes low-utility tokens *during processing*. This directly reduces memory and latency while preserving diagnostically relevant context. Crucially, the separation between retained and discarded tokens provides an implicit saliency signal, mirroring how pathologists ignore uninformative tissue and concentrate on diagnostically meaningful regions. In contrast to scale-heavy pipelines such as Prov-GigaPath and PRISM, our design treats efficiency and interpretability as first-class principles, offering a practical and effective path for vision–language modeling in digital pathology.

3 Method

Our goal is a compute-efficient whole-slide encoder that scales to gigapixel slides while remaining interpretable. To this end, we build SLIM, a three-part architecture: (i) pretrained pathology *patch embeddings* to avoid tile-level pretraining, (ii) a *long-context slide encoder* that preserves global spatial structure, and (iii) *learned token pruning* that shortens sequences during processing without sacrificing task-relevant context. The slide encoder produces a single `[CLS]` representation used for multimodal alignment with frozen PubMedBERT via a CLIP-style contrastive objective; pruning decisions are learned with auxiliary supervision during training and become a lightweight Top-K selector at inference. An overview of the SLIM data flow and the placement of pruning modules is shown in Figure 1. We next detail the dataset and preprocessing, the slide encoder, the pruning mechanism, and the alignment objective.

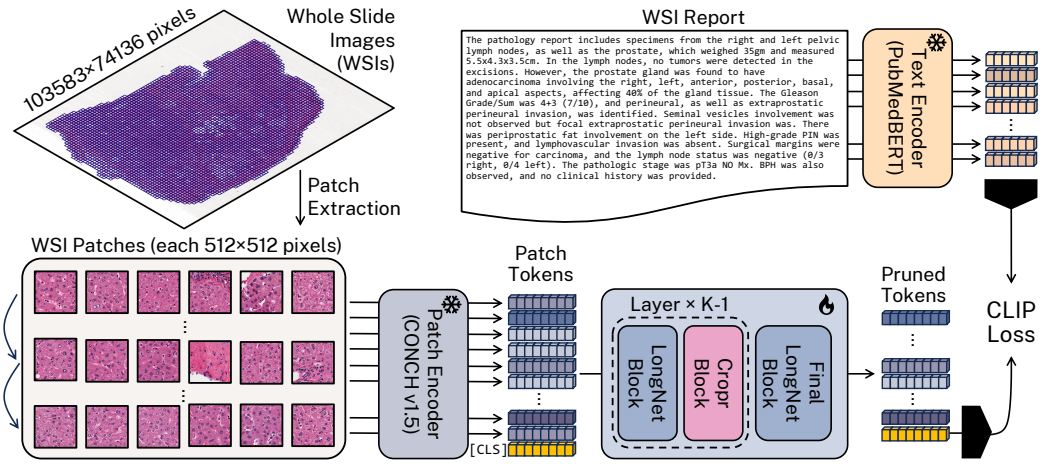

Figure 1: **Overview of our SLIM model.** WSI patch embeddings from CONCH v1.5 are processed by a LongNet-based slide encoder with interleaved Cropr modules. The resulting `[CLS]` embedding is aligned with PubMedBERT report embeddings through a CLIP-style contrastive objective. Details of pruning dynamics are shown in Fig. 2, and the Cropr block design is given in Fig. 3.

3.1 Data and Preprocessing

Dataset curation. Our corpus is constructed from TCGA by combining slide identifiers and reports provided by SlideBench [3] and HistGen [9]. It contains 7,568 WSIs in total, 725 from SlideBench and 6,843 from HistGen. To prevent patient-level information leakage, we retain at most one slide per case, ensuring that each slide corresponds to a unique case ID. We split the dataset into training/validation/test in an 80/10/10 ratio.

Patch extraction and embeddings. WSIs are tiled into non-overlapping 512×512 patches at $20\times$ magnification using TRIDENT [22]. Each patch is encoded with CONCH v1.5 [13], a pathology-specialized encoder trained on large-scale histology data. Using pretrained patch embeddings avoids costly tile-level pretraining while providing strong domain priors.

3.2 Slide Encoder

Whole-slide images contain hundreds of thousands of patch tokens, and simply pooling their embeddings discards spatial dependencies critical for diagnosis. The role of the slide encoder is to aggregate these tokens into a compact sequence while preserving long-range context and spatial structure. This representation must be efficient enough to handle gigapixel slides yet expressive enough to support multimodal alignment with clinical text. To meet these requirements, we adopt a LongNet-based transformer backbone, which scales gracefully to ultra-long sequences, and enrich it with two-dimensional rotary positional encodings that capture the grid structure of WSIs.

Backbone: LongNet. For slide-level modeling, we adopt LongNet [6], which employs dilated attention to efficiently expand receptive fields without quadratic scaling. We prepend a learnable `[CLS]` token whose embedding serves as the global slide descriptor.

2D rotary positional encodings. To encode spatial structure, we use two-dimensional rotary positional encodings (RoPE) [10]. Patch coordinates from TRIDENT are normalized into integer grid indices and mapped into sinusoidal rotations along x and y channels. This yields translation-consistent embeddings while preserving locality.

3.3 Token Pruning

To make ultra-long sequences tractable, we insert Token Cropr [1] modules between transformer blocks. Each module learns a token relevance score during training and prunes low-utility tokens during processing, shrinking sequence length with depth while preserving task-relevant context.

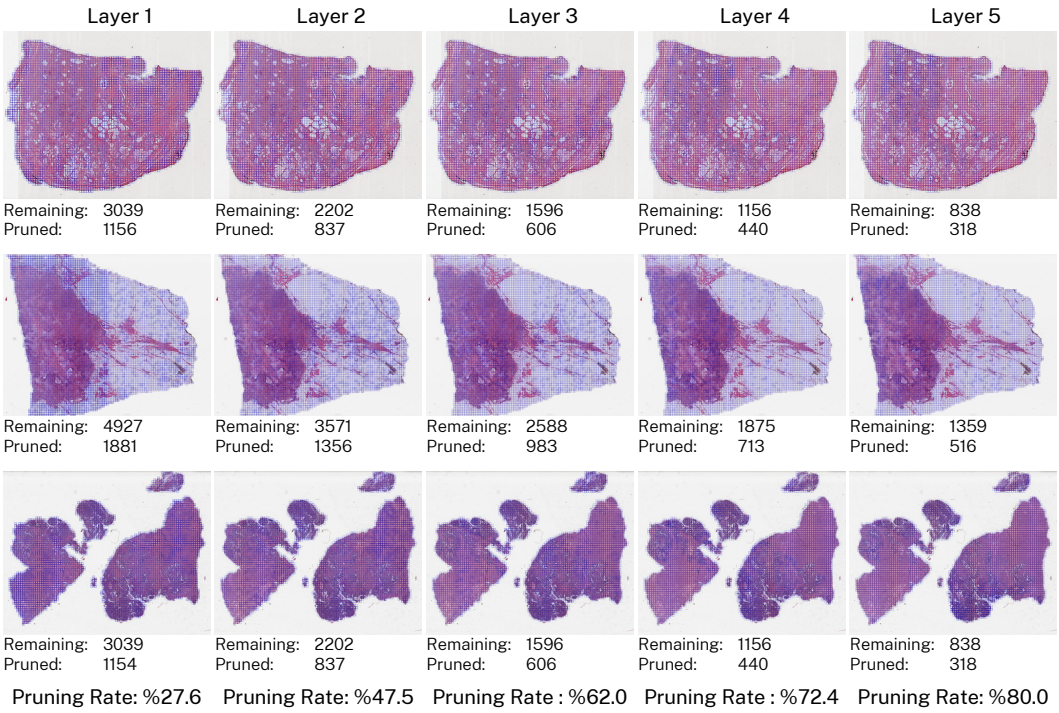

Figure 2: **Pruning across layers.** Retained tokens shrink across depth (light color = pruned), showing relevant regions preserved by Cropr. Best viewed when zoomed-in and in color.

Figure 2 illustrates how the number of retained tokens decreases across layers and how the pruned tokens implicitly highlight diagnostically salient regions.

During training, each Cropr block consists of a router (scorer + Top-K selector), an aggregator, and an auxiliary head for a CLIP loss, providing supervision for token relevance (see Figure 3). At inference, only the lightweight router blocks remain, introducing negligible overhead. This design not only accelerates training and inference but also produces saliency-like signals indicating which regions are considered informative.

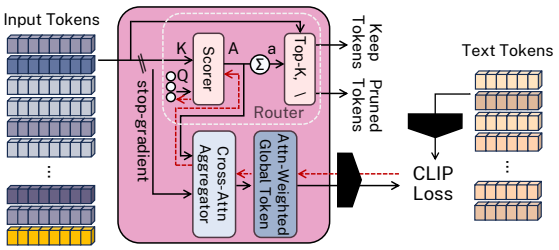

Figure 3: **Cropr module.** Each Cropr module consists of a router, an aggregator, and an auxiliary head. During training, the scorer in the router estimates token salience, the Top-K selector retains the most relevant tokens, and the aggregator with auxiliary head provides supervision to learn pruning decisions through a CLIP-style contrastive loss. At inference, only the router remains, yielding a minimal overhead.

This mechanism also allows for flexible, per-sample control of pruning rates, even at inference, and yields substantial sequence-length reductions without modifying the backbone. Clinically, it mirrors how pathologists work: skimming broad tissue at low magnification and concentrating on diagnostically informative regions. The resulting kept–vs.–pruned token sets serve as implicit saliency maps, exposing which areas the model prioritizes for multimodal alignment and improving interpretability without additional supervision.

Router (Scoring and Selection). Given input tokens $\mathbf{X} \in \mathbb{R}^{M \times D}$, a cross-attention based scorer assigns token relevance:

$$\mathbf{A} = \text{CrossAttn}(\mathbf{Q}, \mathbf{X}), \;\; \mathbf{a}_m = \sum_{n=1}^{N} \mathbf{A}_{nm}. \tag{1}$$

where $\mathbf{Q} \in \mathbb{R}^{N \times D}$ are learnable queries and $\mathbf{A} \in \mathbb{R}^{N \times M}$ are attention weights. The Top-K tokens by score are kept:

$$\mathbf{X}_{\text{keep}} = \text{Top-}K(\mathbf{X} \,|\, \mathbf{a}), \;\; \mathbf{X}_{\text{prune}} = \mathbf{X} \setminus \mathbf{X}_{\text{keep}}. \tag{2}$$

The [CLS] token is always preserved.

Aggregator and Auxiliary Head (Training Only). To provide supervision, Cropr reuses $\mathbf{A}$ to compute attention-weighted features:

$$\mathbf{Z} = \text{softmax}(\frac{\mathbf{A}}{\sqrt{D}})\,\mathbf{X}, \;\; \hat{\mathbf{Z}} = \text{MLP}(\text{LN}(\mathbf{Z})) + \mathbf{Z}, \tag{3}$$

which are fed into an auxiliary prediction head. A stop-gradient ensures Cropr-specific losses do not backpropagate into the main encoder. At inference, both the aggregator and head are discarded.

Inference-Time Simplification. At test time, we avoid constructing the full attention map $\mathbf{A}$. Since $\mathbf{a} = \sum_n \mathbf{Q}_n \mathbf{K}(\mathbf{X})^\top$, we pre-aggregate the queries $\bar{\mathbf{q}} = \sum_n \mathbf{Q}_n$ and compute:

$$\mathbf{a} = \bar{\mathbf{q}}\,\mathbf{K}(\mathbf{X})^\top, \tag{4}$$

where $\mathbf{K}(\mathbf{X})$ are token keys. Thus, each Cropr reduces to a single dot-product plus Top-K selection, as efficient as random pruning but with learned saliency.

Pruning schedule. Cropr modules are inserted after each transformer block except the last. We specify a global keep rate ρ (e.g., $\rho = 0.2$ corresponds to pruning 80% of tokens overall). Let $M^{(0)}$ denote the number of prunable tokens before the first Cropr. Each of the $L-1$ Cropr modules then applies the same per-block keep rate $k = \rho^{1/(L-1)}$, so the sequence length evolves multiplicatively as

$$M^{(\ell+1)} = \lfloor k\, M^{(\ell)} \rceil.$$

After the last Cropr, this yields $M^{(L-1)} \approx \rho\, M^{(0)}$ prunable tokens. The final transformer block does not alter the count; including the always-kept [CLS], the model processes $\approx M^{(L-1)} + 1$ tokens. This schedule distributes pruning evenly across depth while ensuring that the global target keep rate is satisfied. At inference, pruning rates can be reduced further without retraining since the scorer no longer depends on the auxiliary head.

3.4 Multimodal Alignment

We align slides and reports with a CLIP-style contrastive objective [15, 4]. For a batch of size B, slide embeddings $\mathbf{v}_i$ and text embeddings $\mathbf{t}_i$ are normalized and scored $s_{ij} = \langle \mathbf{v}_i, \mathbf{t}_j \rangle / \tau$. The InfoNCE loss is applied symmetrically:

$$\mathcal{L}_{\text{CLIP}} = -\tfrac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(s_{ii})}{\sum_j \exp(s_{ij})} + \log \frac{\exp(s_{ii})}{\sum_j \exp(s_{ji})} \right]. \tag{5}$$

As the text encoder, we use PubMedBERT [8], a domain-specialized biomedical language model. To reduce compute, PubMedBERT is frozen and only a projection layer is trained. On the vision side, the [CLS] token passes through a projection to the same space.

3.5 Implementation Details and Complexity

Our LongNet has six transformer blocks, each with 16 heads, embedding dimension 768, and MLP hidden size 3072. Dilation ratios follow 1,2,4,8,16 for segment lengths up to 16k tokens. The model contains approximately 66 million parameters. Training runs for 20 epochs (2 warmup) with AdamW (lr $= 5 \times 10^{-4}$, weight decay 0.1, cosine schedule). Batch size is 16 with BF16 precision and gradient checkpointing. All LayerNorms, however, are computed in FP32 for numerical stability. Cropr uses one learnable query, single head, and a lightweight MLP auxiliary head (discarded at inference).

Complexity. Let $C_{\text{attn}} \propto \sum_{\ell=1}^{L} M_\ell^2$ be attention cost. With Cropr, M_ℓ shrinks linearly with depth, turning C_{attn} from quadratic in M to a much smaller sum of squares. Because the inference-time Cropr scorer reduces to a single vector–matrix product $\mathbf{a} = \bar{\mathbf{q}}\,\mathbf{K}(\mathbf{X})^\top$ per module (no auxiliary head, no materialized attention), overhead is negligible compared to the savings from smaller M_ℓ.

4 Experimental Results

4.1 Datasets and Evaluation Setup

We evaluate our model on two datasets: the held-out test set of our curated TCGA dataset (Sec. 3.1) and the publicly available EBRAINS dataset [16]. On TCGA, we perform both linear probing and retrieval-based evaluation, while on EBRAINS we restrict to linear probing due to limited availability of paired text annotations.

TCGA test set. For linear probing and zero-shot classification, we follow a setup similar to evaluation on TCGA-OT in TITAN [7], using OncoTree codes to define tumor-type labels. OncoTree codes could not be determined for 10 cases, leaving 747 samples available for classification, spanning 42 diagnostic classes. For linear probing, we discard the classes with less than 10 samples, resulting in a dataset of 630 samples and 21 classes. For zero-shot classification, we use the text prompts and templates from the evaluation on TCGA-OT in TITAN [7]. The zero-shot results are presented in Appendix section A.

EBRAINS dataset. For linear probing on EBRAINS, we apply a minimum class size threshold of 30 to mitigate class imbalance. After filtering, the dataset contains 2,319 slides across 30 diagnostic categories, providing a moderately large-scale benchmark for evaluating cross-dataset generalization.

Baselines. We compare SLIM against two reference models. First, Prov-GigaPath [20], a LongNet-based foundation model trained on 171,189 WSIs with approximately 85M parameters, representing a scale-heavy approach reliant on massive pretraining. Second, a LongNet backbone without Cropr, which shares the same architecture and training protocol as our model but processes all tokens without pruning, referred to as "SLIM (No Pruning)", thereby isolating the effect of token pruning.

Evaluation metrics. For linear probing, we report a broad set of metrics to capture performance beyond overall accuracy. Balanced accuracy accounts for class imbalance, while Cohen's kappa and its normalized-weighted variant (NW Kappa) measure agreement beyond chance. We further include weighted F1 to reflect the trade-off between precision and recall across imbalanced classes, and AUROC to evaluate discrimination quality independent of threshold.

For retrieval experiments on the TCGA dataset, we report both ranking-based and recall-based metrics. The average rank (Img2Text / Text2Img) reflects the mean position of the correct match in the retrieval list, with lower values indicating stronger alignment. Recall at 10 (R@10) measures the proportion of queries where the correct counterpart appears among the top ten retrieved results, providing a thresholded view of retrieval quality.

4.2 Experiments on TCGA Dataset

Table 1 reports linear probing results across multiple classification metrics, while Table 2 summarizes retrieval performance. Several important findings emerge from these experiments. First, SLIM consistently outperforms the scale-heavy Prov-GigaPath baseline [20] across nearly all classification metrics. Despite using far fewer training slides, our framework achieves higher accuracy, balanced accuracy, weighted F1, and AUROC, demonstrating that efficiency-first design can rival or even surpass large-scale pretraining strategies. This highlights the value of token pruning as an alternative to brute-force scaling. Second, comparing our two variants reveals that pruning does not degrade downstream accuracy, in fact, it slightly improves it. While SLIM shows a small drop in Kappa and NW Kappa scores, it achieves higher accuracy, balanced accuracy, and weighted F1 compared to the no-pruning backbone, SLIM (No Pruning). This indicates that pruning not only reduces sequence length and computation but may also act as a form of regularization, focusing the model on diagnostically relevant tokens.

Retrieval experiments in Table 2 show that pruning has minimal effect on cross-modal alignment. SLIM maintains comparable average ranks and recall@10 scores to the no-pruning backbone, despite processing substantially fewer tokens. This suggests that the model can preserve alignment quality while operating at a fraction of the computational cost. Together, these findings demonstrate that token pruning offers a favorable efficiency–accuracy trade-off on TCGA. By reducing token counts

Table 1: **Linear probing results on TCGA.** Token pruning improves accuracy and F1 over the no-pruning baseline and outperforms Prov-GigaPath with far less compute.

Model	Accuracy	Bal. Acc.	Kappa	NW Kappa	Weighted F1	AUROC
SLIM (No Pruning)	0.905	0.919	**0.918**	**0.927**	0.902	**0.998**
SLIM	**0.921**	**0.931**	0.888	0.905	**0.919**	0.996
ProvGigaPath	0.825	0.790	0.617	0.745	0.829	0.973

without sacrificing performance, SLIM enables whole-slide vision–language modeling under realistic compute budgets while retaining strong classification and retrieval ability.

Table 2: **Retrieval results on TCGA.** Cross-modal retrieval performance in image-to-text and text-to-image settings. Pruning maintains retrieval quality while reducing sequence length.

Model	Img2Text Avg. Rank	Text2Img Avg. Rank	Img2Text R@10	Text2Img R@10
SLIM (No Pruning)	32.021	32.338	0.316	0.357
SLIM	31.925	31.563	0.308	0.355

4.3 Experiments on EBRAINS Dataset

We further assess our model on the EBRAINS dataset using linear probing (Table 3). Compared to the TCGA results, performance on EBRAINS is generally lower for our model variants, reflecting the domain shift between datasets. Prov-GigaPath achieves the highest scores overall, benefiting from its large-scale pretraining on hundreds of thousands of WSIs. Nonetheless, our Cropr-enabled SLIM model closely matches the no-pruning backbone across all metrics, with slightly higher accuracy and balanced accuracy despite operating on shorter sequences. This again suggests that pruning can be applied without sacrificing generalization, even under distribution shift. Importantly, while Prov-GigaPath holds an advantage in absolute accuracy on EBRAINS, our model is more lightweight and trained on a much smaller dataset, showing the practical efficiency of our approach.

Table 3: **Linear probing on EBRAINS.** Classification results under domain shift. While Prov-GigaPath achieves the strongest absolute performance, our Cropr-enabled model matches the no-pruning backbone with higher efficiency.

Model	Accuracy	Bal. Acc.	Kappa	NW Kappa	Weighted F1	AUROC
SLIM (No Pruning)	0.707	0.622	0.696	0.695	0.690	0.970
SLIM	0.711	0.654	0.650	0.660	0.693	0.970
ProvGigaPath	**0.802**	**0.756**	**0.737**	**0.756**	**0.797**	**0.984**

The confusion matrices for EBRAINS are provided in the Supplementary. As noted in neuropathology practice, several tumor classes in this dataset (e.g., IDH-mutant versus IDH-wildtype astrocytomas, or oligodendrogliomas defined by 1p/19q co-deletion) are inherently difficult to distinguish on H&E morphology alone without molecular profiling. The observed misclassifications therefore reflect both model limitations and the underlying diagnostic challenges, highlighting the importance of multimodal integration in future work.

4.4 Effect of Pruning Rate

Table 4 compares different pruning rates against the no-pruning baseline. On TCGA, accuracy, balanced accuracy, and F1 steadily increase up to 80% pruning, surpassing the baseline and showing that Cropr not only reduces sequence length but can also improve discriminative performance by removing redundant tokens. In contrast, EBRAINS shows a different pattern: modest pruning around 50% gives small improvements, but heavier pruning reduces performance across most metrics. This difference likely reflects dataset characteristics. TCGA is large and relatively homogeneous, offering more redundancy to prune away, whereas EBRAINS is smaller and more variable, making aggressive pruning riskier. Overall, these results suggest that pruning is not simply a way to save computation; under the right conditions it can also enhance predictive accuracy, though the optimal rate depends

on dataset properties. Importantly, in both datasets pruning consistently highlights diagnostically informative regions, reinforcing its value for both efficiency and interpretability.

Table 4: **Classification Performance at Different Pruning Rates.** Bold highlights the best score per dataset. On TCGA, performance improves steadily up to 80% pruning, surpassing the no-pruning baseline, suggesting that Cropr removes redundant tokens while preserving diagnostic signals. On EBRAINS, moderate pruning offers slight gains over the baseline, but performance declines at higher pruning levels, reflecting dataset-specific sensitivity.

Dataset	Pruning Rate	Accuracy	Bal. Acc.	Kappa	NW Kappa	Weighted F1	AUROC
TCGA	No pruning	0.905	0.919	0.918	0.927	0.902	**0.998**
	50%	0.857	0.868	0.882	0.887	0.854	0.996
	60%	0.857	0.832	0.850	0.867	0.839	0.995
	70%	0.889	0.903	0.851	0.873	0.880	0.996
	80%	**0.921**	**0.931**	**0.888**	**0.905**	**0.919**	0.996
	90%	0.873	0.848	0.887	0.897	0.847	0.997
EBRAINS	No pruning	0.707	0.622	0.696	0.695	0.690	0.970
	50%	**0.716**	0.645	**0.696**	**0.681**	0.700	0.973
	60%	0.694	0.649	0.636	0.649	0.679	0.962
	70%	0.711	0.637	0.611	0.641	**0.701**	**0.974**
	80%	0.711	**0.654**	0.650	0.660	0.693	0.970
	90%	0.677	0.611	0.623	0.636	0.673	0.971

4.5 Efficiency Analysis

We measure efficiency by computing FLOPs on a representative slide with 17,182 patches. The no-pruning baseline requires 2.32T FLOPs, while pruning progressively lowers this cost (Table 5). At a pruning rate of 0.8, FLOPs are halved, and at 0.9 they drop by over 50%. In contrast, Prov-GigaPath consumes 23.1T FLOPs for the same slide–more than 10× higher due to its larger scale. These results show that pruning within the encoder delivers substantial computational savings without degrading retrieval or classification accuracy (Sec. 4.2–4.4), making whole-slide vision–language modeling practical under realistic hardware constraints.[1]

Table 5: **FLOPs per slide.** Pruning progressively reduces computation relative to the no-pruning baseline, while Prov-GigaPath is over 10× more costly due to its larger scale.

Model / Prune Rate	FLOPs	Relative
No pruning	2.32T	1.0×
50% pruning	1.68T	0.72×
60% pruning	1.54T	0.66×
70% pruning	1.38T	0.59×
80% pruning	1.20T	0.52×
90% pruning	0.98T	0.42×
Prov-GigaPath	23.1T	9.9×

5 Conclusions

We introduced SLIM, an efficiency-first framework for whole-slide vision–language modeling that integrates token pruning directly into the slide encoder. By interleaving LongNet blocks with Cropr modules, SLIM shortens sequences while preserving diagnostically relevant content, reducing both memory footprint and FLOPs. Experiments on TCGA and EBRAINS show that pruning preserves, and in some cases even improves, classification and retrieval performance, with outcomes shaped by the interplay of pruning rate, model depth, and dataset characteristics. Efficiency analysis further confirm that our Cropr-based encoder achieves substantial computational savings, over an order of magnitude compared to scale-heavy baselines such as Prov-GigaPath, while remaining competitive in downstream tasks. Together, these findings establish token pruning as a practical design principle for resource-conscious digital pathology models. Looking ahead, SLIM offers a lightweight and interpretable component for multimodal GenAI frameworks that integrate histology, reports, and molecular data into clinically useful systems.

[1]Measured on a 193,224×90,014 WSI: Using 512×512 patches, our encoder produces 17,182 tokens, whereas Prov-GigaPath with 256×256 patches produces 54,023 tokens, contributing to its higher FLOP count.

References

[1] Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Token cropr: Faster vits for quite a few tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9740–9750, June 2025.

[2] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *International Conference on Learning Representations*, 2023.

[3] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5134–5143, June 2025.

[4] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, June 2023.

[5] Mohamed Dhouib, Davide Buscaldi, Sonia Vanier, and Aymen Shabou. Pact: Pruning and clustering-based token reduction for faster visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14582–14592, 2025.

[6] Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens, 2023.

[7] 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, Bowen Chen, Cristina Almagro-Perez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Daisuke Komura, Akihiro Kawabe, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. Multimodal whole slide foundation model for pathology, 2024.

[8] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, October 2021.

[9] Zhengrui Guo, Jiabo Ma, Yingxue Xu, Yihui Wang, Liansheng Wang, and Hao Chen. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 189–199. Springer, 2024.

[10] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer, 2024.

[11] Yuci Liang, Xinheng Lyu, Meidan Ding, Wenting Chen, Jipeng Zhang, Yuexiang Ren, Xiangjian He, Song Wu, Sen Yang, Xiyue Wang, Xiaohan Xing, and Linlin Shen. Wsi-llava: A multimodal large language model for whole slide image, 2024.

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023.

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

[14] Weimin Lyu, Qingqiao Hu, Kehan Qi, Zhan Shi, Wentao Huang, Saumya Gupta, and Chao Chen. Efficient whole slide pathology vqa via token compression, 2025.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[16] T. Roetzer-Pejrimovsky, A. C. Moser, B. Atli, C. C. Vogel, P. A. Mercea, R. Prihoda, E. Gelpi, C. Haberler, R. Höftberger, J. A. Hainfellner, B. Baumann, G. Langs, and A. Woehrer. The digital brain tumour atlas, an open histopathology resource, 2022. Data set.

[17] Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[18] George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D. Kunz, Juan A. Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, Matthew Hanna, Michal Zelechowski, Julian Viret, Neil Tenenholtz, James Hall, Nicolo Fusi, Razik Yousfi, Peter Hamilton, William A. Moye, Eugene Vorontsov, Siqi Liu, and Thomas J. Fuchs. Prism: A multi-modal generative foundation model for slide-level histopathology, 2024.

[19] 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, 2024.

[20] 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*, 2024.

[21] 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*, Aug 2022, 2022.

[22] Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025.