Supplementary Material for 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

This supplementary material provides additional experimental analyzes and supporting details. We begin with zero-shot classification results on the TCGA test set (Sec. A), followed by confusion matrices for TCGA and EBRAINS (Sec. B). We then present ablations on model depth (Sec. C) and positional embeddings (Sec. D), as well as qualitative visualizations of progressive pruning (Sec. E). Finally, we provide the class distributions of the EBRAINS (Sec. F) and TCGA (Sec. G) datasets to support reproducibility and future benchmarking.

A Zero-Shot Classification on the Test Set

We evaluate zero-shot classification on the TCGA test set using OncoTree codes, resulting in a total of 747 samples. Table 1 summarizes the performance across multiple metrics, showing that token pruning improves results consistently over the no-pruning baseline.

Table 1: **Zero-shot classification results on TCGA.** Token pruning improves performance across all reported metrics.

Model	**Accuracy**	**Bal. Acc.**	**Weighted Precision**	**Weighted Recall**	**Weighted F1**	**AUROC**
SLIM (No Pruning)	0.641	0.498	0.692	0.641	0.605	0.981
SLIM	**0.656**	**0.531**	**0.693**	**0.655**	**0.609**	**0.982**

To generate predictions, we adopt the text prompts introduced in [1], which provide diverse natural language descriptions of each diagnostic class. The full prompt set is listed below:

- CLASSNAME.
- an image of CLASSNAME.
- the image shows CLASSNAME.
- the image displays CLASSNAME.
- the image exhibits CLASSNAME.
- an example of CLASSNAME.
- CLASSNAME is shown.
- this is CLASSNAME.
- I observe CLASSNAME.
- the pathology image shows CLASSNAME.
- a pathology image shows CLASSNAME.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

- the pathology slideshows CLASSNAME.
- shows CLASSNAME.
- contains CLASSNAME.
- presence of CLASSNAME.
- CLASSNAME is present.
- CLASSNAME is observed.
- the pathology image reveals CLASSNAME.
- a microscopic image of showing CLASSNAME.
- histology shows CLASSNAME.
- CLASSNAME can be seen.
- the tissue shows CLASSNAME.
- CLASSNAME is identified.

For each class, embeddings are computed for all prompts and aggregated via average pooling, using the class list in Table 5 as the label space. The resulting confusion matrix is provided in Figure 1, highlighting class-specific strengths and weaknesses.

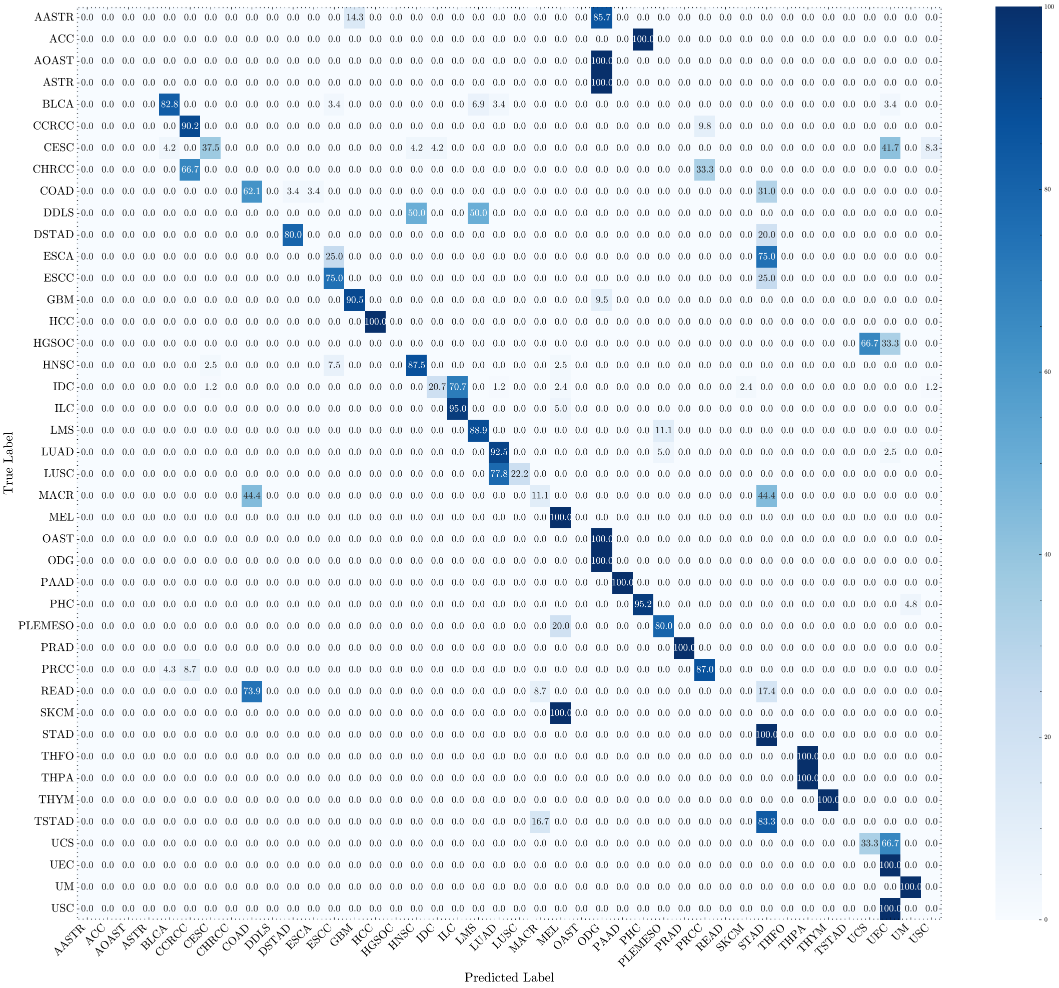
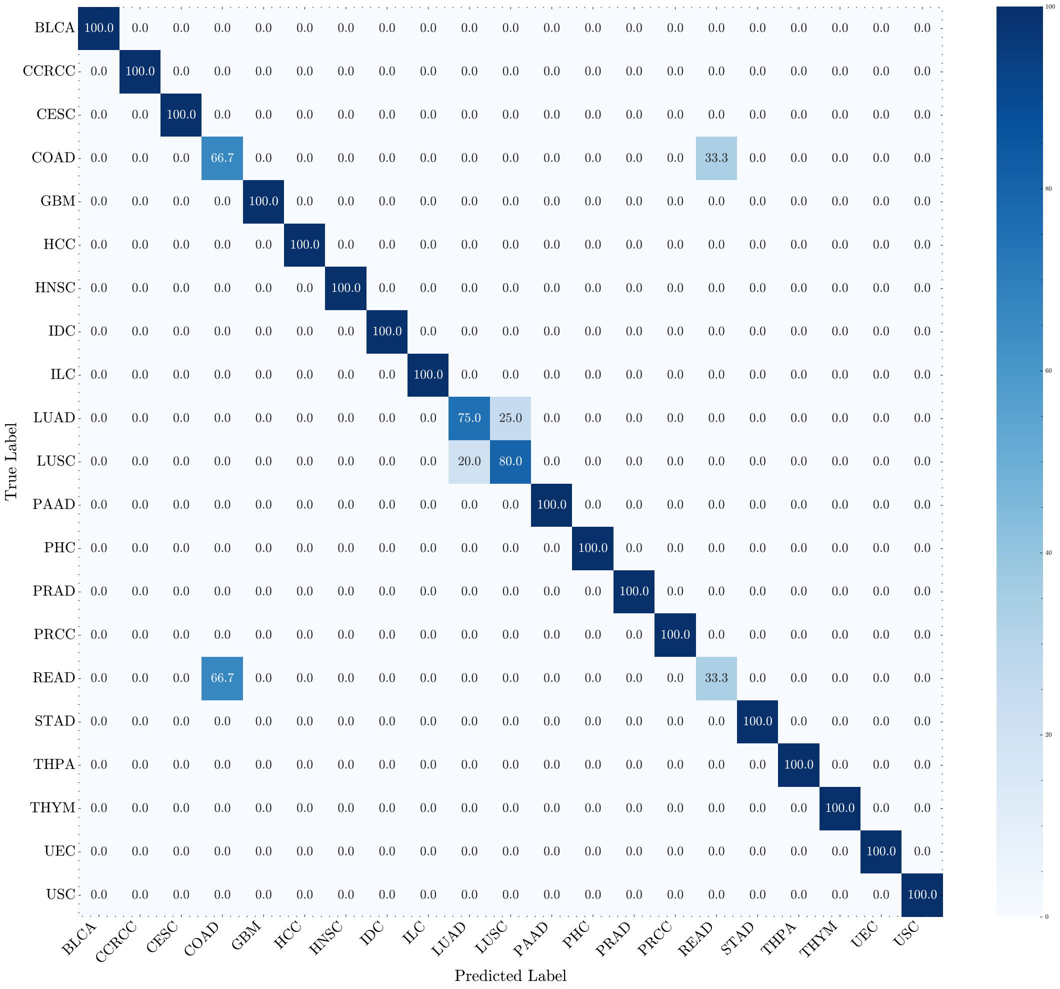


Figure 1: **Confusion matrix illustrating zero-shot classification performance of our model on the TCGA test set.**

B Confusion Matrices for Linear Probing

Figures 2 and 3 show confusion matrices for linear probing on the TCGA and EBRAINS datasets, respectively. These visualizations provide a detailed view of class-level performance beyond the aggregate metrics reported in the main text.

On TCGA, the confusion matrix highlights that most tumor classes are well separated, with errors concentrated among morphologically similar subtypes. On EBRAINS, misclassifications are more frequent, reflecting the smaller dataset size, greater heterogeneity, and cross-domain shift. In particular, several tumor types—such as IDH-mutant versus IDH-wildtype astrocytomas or oligodendrogliomas defined by 1p/19q co-deletion—are nearly indistinguishable from H&E morphology alone without molecular profiling. These patterns are consistent with the known diagnostic challenges in neuropathology, where integrated histological and molecular analysis is required.



	BLCA	CCRCC	CESC	COAD	GBM	HCC	HNSC	IDC	ILC	LUAD	LUSC	PAAD	PHC	PRAD	PRCC	READ	STAD	THPA	THYM	UEC	USC
BLCA	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CCRCC	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CESC	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
COAD	0.0	0.0	0.0	66.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	33.3	0.0	0.0	0.0	0.0	0.0
GBM	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HCC	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HNSC	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
IDC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ILC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LUAD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	75.0	25.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LUSC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	80.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PAAD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PHC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PRAD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PRCC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
READ	0.0	0.0	0.0	66.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	33.3	0.0	0.0	0.0	0.0	0.0
STAD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
THPA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
THYM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
UEC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0
USC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0

Figure 2: **Confusion matrix illustrating linear probing performance of our model on the TCGA test set.**

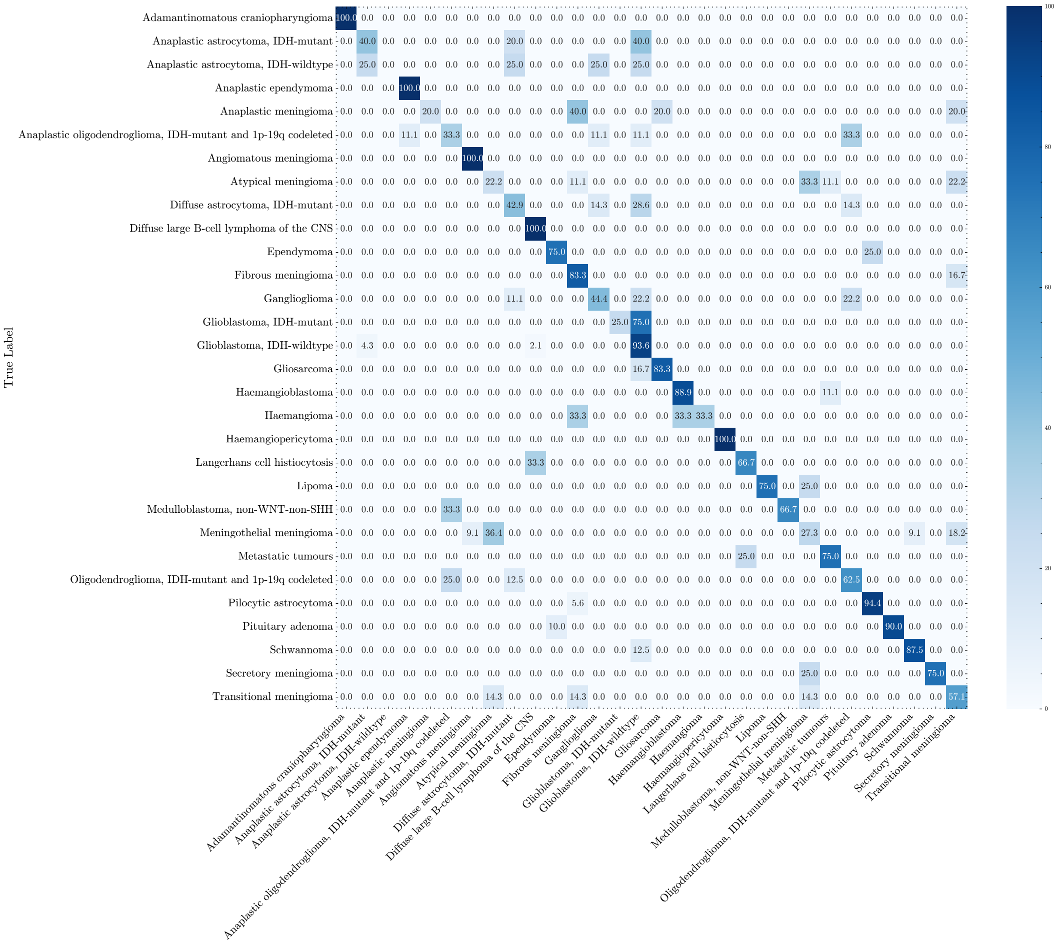


Figure 3: **Confusion matrix illustrating linear probing performance of our model on the EBRAINS dataset.**

C Effect of Model Depth

We analyze the impact of transformer depth on classification performance, with results summarized in Table 2. Both TCGA and EBRAINS exhibit an optimal range, with six layers providing the best balance of accuracy, balanced accuracy, and F1. This depth appears sufficient to capture global context without introducing unnecessary complexity. Adding more layers does not consistently improve accuracy, and on TCGA performance often plateaus or slightly degrades, indicating diminishing returns once model capacity exceeds dataset scale. On EBRAINS, a smaller and more heterogeneous dataset, depth has a stronger effect: shallow models (4–5 layers) underperform, while deeper configurations beyond six trade off modest drops in accuracy for gains in agreement metrics such as kappa and AUROC. This suggests that while six layers offer a strong efficiency–accuracy trade-off across datasets, additional depth may improve robustness under variability in acquisition conditions.

Table 2: **Effect of model depth on classification performance.** Results across transformer depths on TCGA and EBRAINS. On both datasets, 6 layers provide the best overall balance of accuracy, agreement, and AUROC, though EBRAINS shows larger gains from depth.

Dataset	Layers	Accuracy	Bal. Acc.	Kappa	Weighted F1	AUROC
TCGA	4	0.841	0.818	0.887	0.810	0.997
	5	0.857	0.880	0.887	0.841	0.995
	6	**0.921**	**0.931**	0.888	**0.919**	0.996
	7	0.889	0.917	0.888	0.888	0.997
	8	0.857	0.892	0.882	0.857	0.994
	10	0.905	0.926	**0.925**	0.901	**0.997**
EBRAINS	4	0.690	0.612	0.611	0.666	0.969
	5	0.685	0.606	0.641	0.664	0.966
	6	**0.711**	**0.654**	0.650	**0.693**	0.970
	7	0.698	0.633	0.611	0.688	**0.971**
	8	0.690	0.611	**0.685**	0.683	0.971
	10	0.681	0.623	0.641	0.666	0.967

D Effect of Positional Embeddings

We compare rotary positional embeddings (RoPE) with standard sine–cosine encodings for slide-level modeling (Table 3). On TCGA, RoPE is consistently stronger across metrics. It improves accuracy by nearly 8 points (0.921 vs. 0.841) and yields stronger agreement metrics and F1. On EBRAINS, RoPE achieves higher accuracy (0.711 vs. 0.681) and balanced accuracy (0.654 vs. 0.599), while maintaining similar AUROC. These results suggest that rotary embeddings are better suited for capturing fine-grained spatial structure in WSIs, particularly under domain shift, where sine–cosine encodings appear less robust.

Table 3: **Effect of positional embeddings.** RoPE yields clear gains on TCGA and improves accuracy/recall-oriented metrics on EBRAINS.

Dataset	Positional Embedding	Accuracy	Bal. Acc.	Kappa	Weighted F1	AUROC
TCGA	RoPE	**0.921**	**0.931**	**0.888**	**0.919**	**0.996**
	SineCos	0.841	0.858	0.805	0.839	0.994
EBRAINS	RoPE	**0.711**	**0.654**	0.650	**0.693**	**0.970**
	SineCos	0.681	0.599	**0.687**	0.666	0.969

E Visualization of Progressive Token Pruning

Figure 4 illustrates how the Cropr modules gradually reduces sequence length across transformer layers. Beyond showing retained and discarded tokens, the visualizations highlight the dynamics of

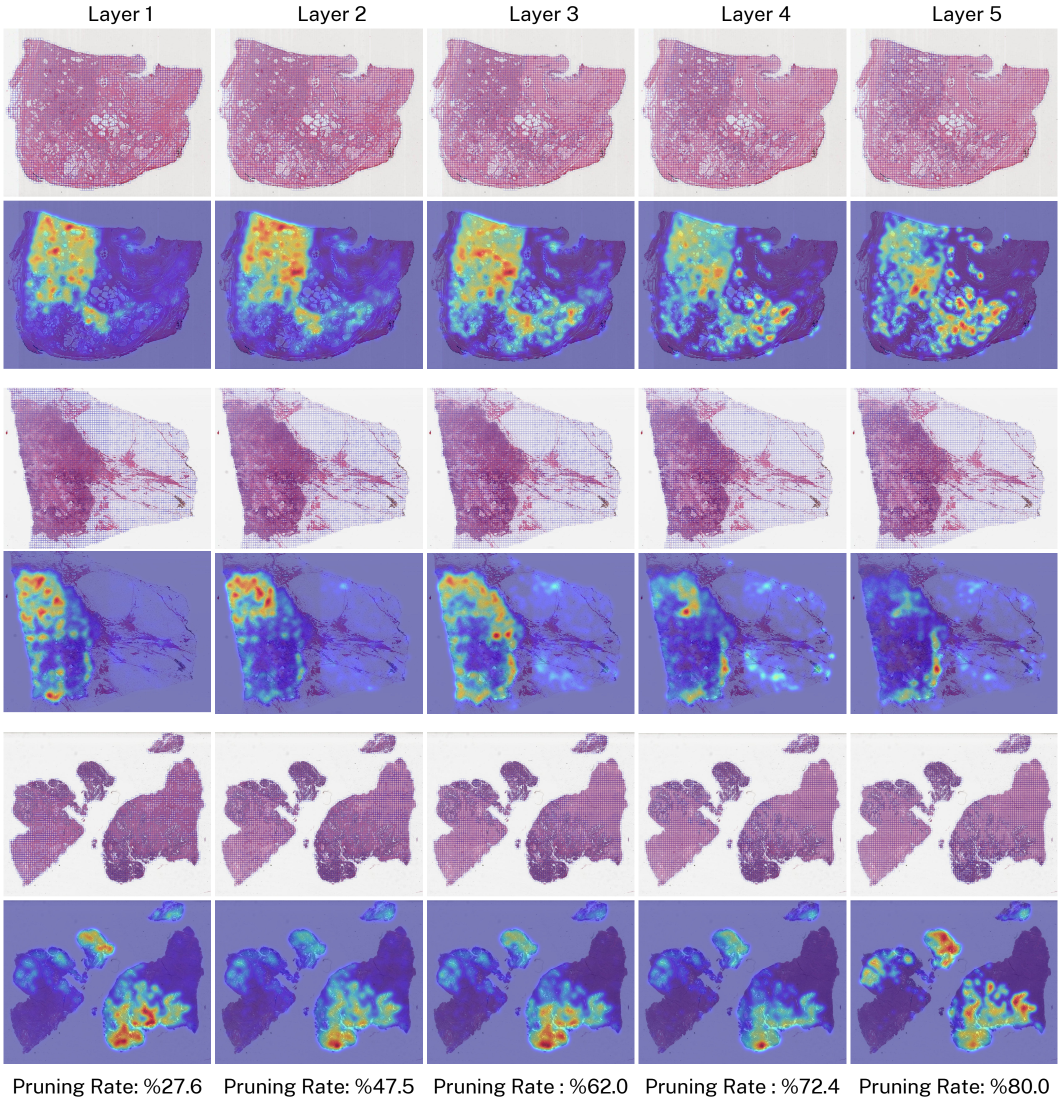


Figure 4: **Visualization of Progressive Token Pruning Across Layers.** Top rows show the retained histology patches (light color), while bottom rows overlay Cropr attention maps, where warmer colors indicate tokens assigned higher salience. As depth increases, uninformative background regions are gradually discarded and the model focuses more tightly on diagnostically relevant tissue areas.

the pruning process itself. The top row of each example shows the remaining tissue patches after pruning, while the bottom row overlays token salience scores derived from Cropr's attention weights.

Early layers eliminate broad, homogeneous regions, while later layers focus the model's capacity on localized tissue structures. The progression reveals that pruning operates hierarchically: it first removes obvious background, then progressively sharpens the representation around diagnostically relevant areas.

This behavior is important for two reasons. First, it shows that token pruning is not a blunt reduction mechanism but an adaptive process that reallocates capacity as depth increases. Second, it provides a form of interpretability: the regions that consistently survive pruning align with areas a pathologist would attend to, suggesting that the model's efficiency gains are coupled with meaningful inductive biases.

F EBRAINS Class Distribution

To support reproducibility, we report the class distribution of the EBRAINS dataset used in our experiments. As described in the main paper, we filtered out categories with fewer than 30 samples, resulting in 2,319 slides spanning 30 diagnostic classes. Table 4 lists the number of samples per class, which may serve as a reference for future benchmarking and comparisons.

Table 4: **EBRAINS class distribution.** Number of samples per class after filtering out categories with fewer than 30 slides. The final dataset contains 2,319 slides across 30 diagnostic classes.

Class	Sample Count
Adamantinomatous craniopharyngioma	85
Anaplastic astrocytoma, IDH-mutant	47
Anaplastic astrocytoma, IDH-wildtype	47
Anaplastic ependymoma	50
Anaplastic meningioma	46
Anaplastic oligodendroglioma, IDH-mutant and 1p-19q codeleted	91
Angiomatous meningioma	31
Atypical meningioma	83
Diffuse astrocytoma, IDH-mutant	70
Diffuse large B-cell lymphoma of the CNS	59
Ependymoma	46
Fibrous meningioma	57
Ganglioglioma	88
Glioblastoma, IDH-mutant	34
Glioblastoma, IDH-wildtype	474
Gliosarcoma	59
Haemangioblastoma	88
Haemangioma	30
Haemangiopericytoma	34
Langerhans cell histiocytosis	32
Lipoma	38
Medulloblastoma, non-WNT-non-SHH	32
Meningothelial meningioma	104
Metastatic tumours	47
Oligodendroglioma, IDH-mutant and 1p-19q codeleted	85
Pilocytic astrocytoma	173
Pituitary adenoma	99
Schwannoma	81
Secretory meningioma	41
Transitional meningioma	68

G TCGA Class Distribution

To support reproducibility, we also provide the distribution of classes in the curated TCGA dataset used in our experiments. After removing slides without valid OncoTree codes, the test set contains 747 slides across 42 diagnostic classes. Table 5 reports the number of samples per class. The prompts are taken from the evaluation setup in [1].

Table 5: **TCGA class distribution.** Number of samples per class in the curated TCGA dataset used in our experiments. The test set contains 747 slides spanning 42 diagnostic classes.

Class	Sample Count	Prompts
IDC	82	invasive ductal carcinoma breast invasive ductal carcinoma ductal carcinoma, no special type IDC
LUSC	45	squamous cell carcinoma lung squamous cell carcinoma squamous carcinoma of the lung LUSC
CCRCC	41	clear cell carcinoma renal clear cell carcinoma clear cell renal cell carcinoma clear cell RCC CCRCC
LUAD	40	lung adenocarcinoma adenocarcinoma of the lung pulmonary adenocarcinoma peripheral lung adenocarcinoma LUAD
HNSC	40	head and neck squamous cell carcinoma HNSCC head and neck SCC oropharyngeal squamous cell carcinoma laryngeal squamous cell carcinoma hypopharyngeal squamous cell carcinoma nasopharyngeal squamous cell carcinoma oral cavity squamous cell carcinoma HNSC
HCC	38	hepatocellular carcinoma liver cancer hepatoma HCC
UEC	35	endometrioid carcinoma uterine endometrioid carcinoma endometrial endometrioid carcinoma endometrial carcinoma, endometrioid type UEC
THPA	33	papillary thyroid carcinoma papillary thyroid cancer papillary thyroid neoplasm PTC thyroid papillary carcinoma
PRAD	33	prostate adenocarcinoma adenocarcinoma of the prostate prostatic adenocarcinoma PRAD
COAD	29	colon adenocarcinoma adenocarcinoma of the colon colorectal adenocarcinoma COAD

Continued on next page

Table 5 – continued from previous page

Class	Sample Count	Prompts
BLCA	29	urothelial carcinoma bladder urothelial carcinoma transitional cell carcinoma bladder cancer BLCA
CESC	24	squamous cell carcinoma of the cervix cervical squamous cell carcinoma cervical SCC CESC
PRCC	23	papillary renal cell carcinoma renal cell carcinoma, papillary type papillary RCC PRCC
READ	23	rectal adenocarcinoma adenocarcinoma of the rectum rectal cancer READ
PHC	21	pheochromocytoma chromaffin cell tumor adrenal pheochromocytoma paraganglioma PHC
GBM	21	glioblastoma multiforme glioblastoma GBM multiforme glioblastoma grade IV astrocytoma
ILC	20	invasive lobular carcinoma breast invasive lobular carcinoma invasive lobular carcinoma of the breast lobular carcinoma breast ILC
PAAD	19	pancreatic adenocarcinoma adenocarcinoma of the pancreas pancreas adenocarcinoma PAAD
STAD	13	diffuse type stomach adenocarcinoma diffuse gastric adenocarcinoma diffuse-type gastric cancer gastric adenocarcinoma, diffuse type STAD
THYM	11	thymoma thymic tumor thymic carcinoma thymic epithelial neoplasm THYM
USC	10	uterine serous carcinoma uterine papillary serous carcinoma serous papillary carcinoma of the uterus serous carcinoma USC

Continued on next page

Table 5 – continued from previous page

Class	Sample Count	Prompts
LMS	9	leiomyosarcoma smooth muscle sarcoma leiomyosarcoma of the soft tissue LMS
ODG	9	oligodendroglioma oligodendroglial tumor oligodendroglioma grade II oligodendroglioma grade III ODG
CHRCC	9	chromophobe renal cell carcinoma chromophobe RCC chromophobe carcinoma renal cell carcinoma, chromophobe type CHRCC
UCS	9	uterine carcinosarcoma uterine malignant mixed Müllerian tumor malignant mixed Müllerian tumor carcinosarcoma of the uterus UCS
MACR	9	mucinous adenocarcinoma mucinous adenocarcinoma of the colon mucinous adenocarcinoma of the rectum mucinous colorectal adenocarcinoma MACR
THFO	9	follicular thyroid carcinoma follicular thyroid cancer follicular carcinoma of the thyroid thyroid follicular cancer THFO
AASTR	7	anaplastic astrocytoma astrocytoma, anaplastic grade III astrocytoma AASTR
UM	6	uveal melanoma uvea melanoma ocular melanoma choroidal melanoma ciliary body melanoma iris melanoma UM
OAST	6	oligoastrocytoma oligoastrocytic tumor oligodendroglial-astrocytic tumor OAST
TSTAD	6	tubular stomach adenocarcinoma tubular adenocarcinoma of the stomach stomach tubular adenocarcinoma gastric tubular adenocarcinoma TSTAD
DSTAD	5	diffuse type adenocarcinoma of the stomach diffuse stomach adenocarcinoma diffuse gastric adenocarcinoma stomach adenocarcinoma, diffuse type DSTAD

Continued on next page

Table 5 – continued from previous page

Class	Sample Count	Prompts
ASTR	5	astrocytoma diffuse astrocytoma fibrillary astrocytoma anaplastic astrocytoma grade II astrocytoma grade III astrocytoma ASTR
PLEMESO	5	pleural mesothelioma epithelioid mesothelioma epithelioid pleural mesothelioma pleural mesothelioma, epithelioid type PLEMESO
MEL	4	acral melanoma melanoma acral lentiginous melanoma MEL
ESCA	4	esophageal adenocarcinoma adenocarcinoma of the esophagus esophageal cancer, adenocarcinoma type ESCA
ESCC	4	esophageal squamous cell carcinoma esophageal SCC squamous cell carcinoma of the esophagus SCC of the esophagus ESCC
SKCM	3	cutaneous melanoma skin melanoma melanoma of the skin malignant melanoma SKCM
HGSOC	3	high-grade serous ovarian cancer high-grade serous carcinoma HGSOC serous papillary carcinoma serous ovarian carcinoma
AOAST	2	anaplastic oligoastrocytoma oligoastrocytoma anaplastic mixed glioma AOAST
DDLS	2	dedifferentiated liposarcoma dedifferentiated liposarcoma variant liposarcoma, dedifferentiated DDLS
ACC	1	adrenocortical carcinoma adrenal cortical carcinoma adrenal cortex carcinoma ACC

References

[1] 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.