DTC-WSI: Dynamic Token Compression for Whole-Slide Images

Tawsifur Rahman[1] ARAHMA34@JHU.EDU
[1] *Biomedical Engineering, Johns Hopkins University*
Aliasghar Tarkhan[2] TARKHAN.ALIASGHAR@GMAIL.COM
[2] *Johnson & Johnson, MedTech*
Rama Chellappa[3] RCHELLA4@JHU.EDU
[3] *Johns Hopkins University*
Alexander S. Baras[4] BARAS@JHMI.EDU
[4] *School of Medicine, Johns Hopkins University*

Editors: Under Review for MIDL 2026

Abstract

Whole-slide images (WSIs) contain tens of thousands of heterogeneous patches, making transformer-based multiple-instance learning (MIL) computationally expensive due to quadratic attention costs and substantial redundancy in tissue morphology. Existing token-reduction approaches for WSI analysis rely primarily on pruning, which discards information early in training and destabilizes optimization under weak supervision. We propose **Dynamic Token Compression for Whole-Slide Images (DTC-WSI)**, a token-efficient MIL framework that performs *progressive*, *importance-aware* WSI compression. DTC-WSI integrates a lightweight saliency network with a multi-stage token compressor that combines *bipartite similarity matching* and *soft differentiable pruning* to gradually eliminate redundant or non-diagnostic patches. During training, soft gates enable stable gradient flow, while inference employs deterministic compression for substantial acceleration. This curriculum-style compression preserves discriminative morphology and dramatically reduces computational burden. Across four WSI benchmarks (TCGA-NSCLC, TCGA-BRCA, TCGA-RCC, PANDA), DTC-WSI achieves **5–10× token reduction**, **up to 5.3× faster inference**, and **20–40% lower memory usage**, while improving MIL classification accuracy by **2–4%** over state-of-the-art baselines. Our results demonstrate that dynamic token compression is a powerful and scalable alternative to pruning, enabling efficient transformer-based WSI analysis while improving accuracy.

Keywords: Computational pathology, Token merging, Dynamic token pruning, Weakly supervised learning

1. Introduction

Whole-slide images (WSIs) are gigapixel-scale pathology scans containing rich and heterogeneous morphological patterns across extremely large spatial extents (1; 2; 3). Because WSIs cannot be processed directly at native resolution, modern computational pathology workflows tile each slide into thousands of fixed-size patches and use multiple instance learning (MIL) to aggregate patch embeddings into slide-level predictions (4; 5; 6). Attention-based MIL approaches such as ABMIL (7), CLAM (8), TransMIL (9), DSMIL (10), and hierarchical models such as HIPT (11) have achieved strong performance on cancer subtype

classification, grading, and prognosis tasks. However, these methods face a critical scalability bottleneck: a single diagnostic WSI often produces tens of thousands of tokens, causing transformer attention layers and MIL scoring modules to incur substantial computational and memory overhead (12; 13; 14).

This challenge is exacerbated by the structural properties of histopathology images. Large regions contain visually redundant or weakly discriminative tissue—including stroma, adipose, necrosis, and repeated tumor textures (15). Treating all patches as independent tokens forces models to process extensive redundancy, increasing computation without adding discriminative signal (16). Prior attempts to mitigate this include hierarchical MIL (17), patch clustering (18), and token pruning (19; 20). Yet pruning irreversibly discards tokens and risks removing diagnostically relevant regions, a severe limitation under weak supervision where slide-level labels provide no guidance for early-stage pruning decisions (21; 22; 23).

Meanwhile, the natural-image community has demonstrated that *token merging* can accelerate Vision Transformers by fusing redundant tokens rather than removing them. Methods such as ToMe (24) exploit similarity structure to merge tokens without losing information. However, these methods have not been adapted to computational pathology, where redundancy patterns are more complex, token counts are orders of magnitude larger, and merging must be guided by task-driven saliency to avoid collapsing diagnostically meaningful structures (25; 26; 27).

To address these limitations, we propose **Dynamic Token Compression for Whole-Slide Images (DTC-WSI)**, a unified framework that integrates *similarity-guided token merging* with *importance-guided pruning* within a progressive, multi-stage compression pipeline. Unlike pruning-only approaches, DTC-WSI preserves diagnostic content by fusing redundant patches into compact representations using efficient bipartite matching. Unlike ToMe-style merging alone, our method incorporates a differentiable *importance network* that learns patch saliency from slide-level supervision and guides compression decisions throughout training. A key innovation of DTC-WSI is its *multi-stage compression strategy*, which gradually reduces token count across several stages rather than performing aggressive reduction at once. This curriculum-like design prevents early information collapse and allows the importance network to stabilize its saliency estimates before significant compression occurs. Through this hybrid design, DTC-WSI addresses a central scalability challenge in computational pathology: how to process gigapixel WSIs efficiently while retaining diagnostically essential information. Our contributions are summarized as follows:

1. **A unified multi-stage token compression framework** that jointly performs similarity-guided token merging and importance-guided pruning, enabling aggressive token reduction while preserving diagnostic morphology.

2. **A differentiable importance network** that learns patch saliency under weak supervision, guiding compression during training and enabling deterministic, high-efficiency inference.

3. **Comprehensive evaluation on four major WSI benchmarks** (TCGA-NSCLC, TCGA-BRCA, TCGA-RCC, PANDA), demonstrating that DTC-WSI achieves **5–10× token reduction**, **up to 5.3× faster inference**, **20–40% lower memory**

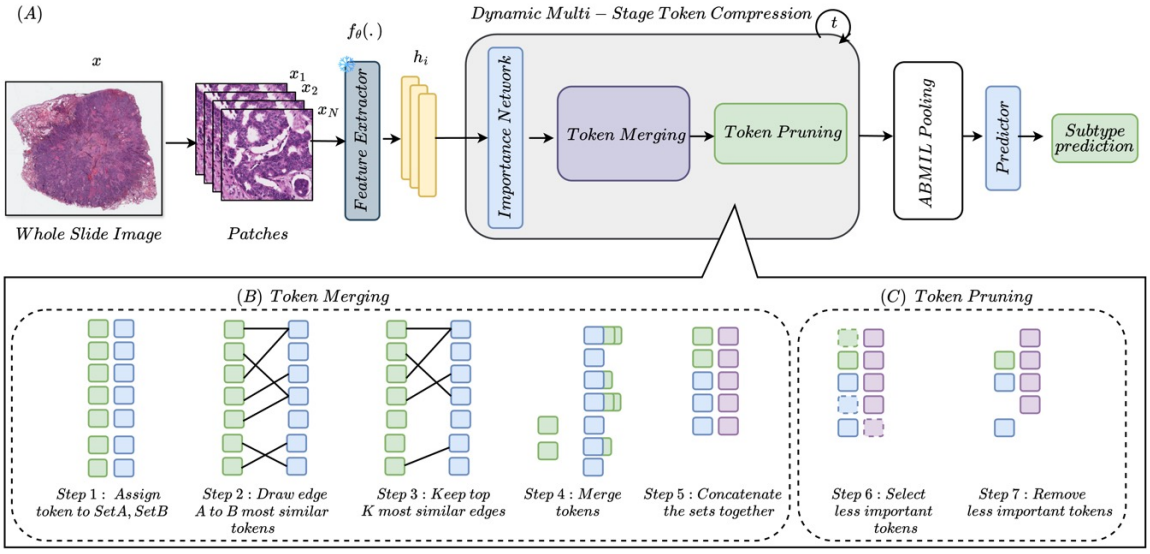

Figure 1: Overview of the proposed **Dynamic Token Compression (DTC-WSI)** framework. **(A)** End-to-end pipeline: patch extraction, feature encoding, multi-stage token compression, and MIL prediction. **(B)** Token merging: similar patches are fused into unified representations via bipartite soft matching. **(C)** Token pruning: low-importance tokens are removed to produce a compact, discriminative set for classification.

usage, and **2–4% accuracy gains** over state-of-the-art MIL and token-efficient baselines.

2. Methods

2.1. Overview

Whole-slide images (WSIs) contain tens of thousands of patches, making conventional MIL and transformer models computationally prohibitive due to quadratic attention and high memory demands. To address this, we propose **Dynamic Token Compression for Whole-Slide Images (DTC-WSI)**, a framework that *learns* to compress WSI patch embeddings in a task-aware manner. DTC-WSI progressively reduces tokens across multiple stages by combining (1) **similarity-guided merging** to fuse redundant patches and (2) **importance-guided pruning** to discard low-saliency regions. Compression is applied *softly* during training, enabling the importance network to learn reliable saliency estimates, and *deterministically* at inference for fast, scalable deployment. The final compact token set is aggregated using attention-based MIL to produce slide-level predictions.

2.2. Patch Extraction and Feature Encoding

A whole-slide image (WSI) is denoted by x, which is tiled into N non-overlapping patches $\{x_1, x_2, \ldots, x_N\}$ after tissue detection and background removal. Each patch is processed by a pretrained encoder (e.g., CONCH (28)) to obtain a semantic feature embedding:

$$h_i^{(0)} = f_\theta(x_i) \in \mathbb{R}^D, \qquad i = 1, \ldots, N, \tag{1}$$

forming the initial token matrix $H^{(0)} = [h_1^{(0)}, \ldots, h_N^{(0)}]^\top \in \mathbb{R}^{N \times D}$.

To preserve spatial information, optional positional embeddings p_i concatenated with the visual features:

$$\tilde{h}_i^{(0)} = [h_i^{(0)} \,\|\, p_i].$$

All supervision is provided at the slide level; no patch-level labels are used during training.

2.3. Importance Network

WSIs contain large regions of redundant or clinically irrelevant tissue, making it essential to quantify which patch embeddings contribute meaningfully to the slide prediction. The **Importance Network** g_ϕ assigns a saliency score to each token at stage t:

$$s_i^{(t)} = g_\phi(\tilde{h}_i^{(t-1)}), \tag{2}$$

where g_ϕ is a two-layer MLP with GELU activation. Scores are normalized into importance weights:

$$\alpha_i^{(t)} = \frac{\exp(s_i^{(t)})}{\sum_{j=1}^{N^{(t)}} \exp(s_j^{(t)})},$$

which induces soft competition among tokens. Early in training, the distribution remains diffuse; as learning progresses, high-saliency tumor regions receive larger weights.

2.4. Dynamic Multi-Stage Token Compression

To prevent catastrophic loss of diagnostic evidence, we adopt a **multi-stage compression** schedule:

$$N^{(0)} = N \;\rightarrow\; N^{(1)} \;\rightarrow\; N^{(2)} \;\rightarrow\; N^{(3)}, \tag{3}$$

where each N_{t+1} is determined by a retention ratio $r : N^{(t+1)} = r \cdot N^{(t)}$

The number of token merges required in stage t is: $K^{(t)} = N^{(t)} - N^{(t+1)}$

Each stage consists of: 1) **Bipartite soft matching for token fusion**, and 2) **Importance-guided pruning**.

2.4.1. Bipartite Soft Matching for Token Fusion

To avoid the $O(N^2)$ complexity of full similarity search, tokens in stage t are partitioned into alternating subsets:

$$A = [\tilde{h}_1^{(t-1)}, \tilde{h}_3^{(t-1)}, \tilde{h}_5^{(t-1)}, \ldots], \quad B = [\tilde{h}_2^{(t-1)}, \tilde{h}_4^{(t-1)}, \tilde{h}_6^{(t-1)}, \ldots].$$

For each aligned pair (i, j), cosine similarity is computed:

$$\mathrm{sim}(i,j) = \frac{\langle \tilde{h}_i^{(t-1)}, \tilde{h}_j^{(t-1)} \rangle}{\|\tilde{h}_i^{(t-1)}\| \, \|\tilde{h}_j^{(t-1)}\|}.$$

A merge utility incorporating importance consistency is defined as:

$$u_{ij}^{(t-1)} = \lambda \, \mathrm{sim}(i,j) - (1-\lambda) \, |\alpha_i^{(t)} - \alpha_j^{(t)}|.$$

The Top-$K^{(t)}$ pairs are fused using importance-weighted averaging:

$$\tilde{h}_l^{(t)} = \frac{\alpha_i^{(t)} \tilde{h}_i^{(t-1)} + \alpha_j^{(t)} \tilde{h}_j^{(t-1)}}{\alpha_i^{(t)} + \alpha_j^{(t)}}. \tag{4}$$

2.4.2. Importance-Guided Token Pruning

After token merging, low-saliency tokens are suppressed. During training, pruning is differentiable:

$$m_l^{(t)} = \sigma\Big(\gamma(\alpha_l^{(t)} - \tau)\Big), \quad \tilde{h}_l^{(t)} = m_l^{(t)} \, \tilde{h}_l^{(t)}.$$

During inference, deterministic Top-$N^{(t)}$ pruning is applied:

$$H^{(t)} = \mathrm{TopK}\Big(H^{(t-1)}, \alpha^{(t-1)}, N^{(t-1)}\Big).$$

2.5. MIL Aggregation and Prediction

After t compression stages, the final tokens $H^{(t)}$ are passed to an attention-based MIL module. Attention weights:

$$a_i = \frac{\exp(w^\top \tanh(W h_i^{(t)}))}{\sum_{j=1}^{M} \exp(w^\top \tanh(W h_j^{(t)}))}.$$

The final slide-level representation is computed as a weighted sum of the compressed tokens, $z = \sum_{i=1}^{N^{(t)}} a_i \tilde{h}_i^{(t)}$, where the attention weights emphasize diagnostically informative regions. This embedding is then passed through a linear classifier followed by a softmax layer to produce the slide-level prediction, $\hat{y} = \mathrm{softmax}(W_c z + b_c)$.

2.6. Loss Function

We supervise slide-level predictions using cross-entropy, $\mathcal{L}_{\mathrm{cls}} = \mathrm{CE}(y, \hat{y})$, and encourage the importance network to assign sparse, selective saliency through an ℓ_1 regularizer, $\mathcal{L}_{\mathrm{sparse}} = \beta \sum_{t=0}^{T} \|\alpha^{(t)}\|_1$. The full training objective combines both terms to promote discriminative yet compact representations, the composite loss is given as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{sparse}}.$$

A complete step-by-step description of the algorithm is provided in **Appendix D**.

3. Results

3.1. Datasets

We evaluated DTC-WSI across four large-scale histopathology classification benchmarks and one cellular-level morphology task to assess both its robustness on diverse cancer subtyping problems and its generalizability beyond WSIs.

TCGA-NSCLC. (29) This dataset comprises 993 whole-slide images (WSIs) from Formalin-Fixed Paraffin-Embedded (FFPE) tissue samples, with 507 slides corresponding to lung adenocarcinoma (LUAD) and 486 to lung squamous cell carcinoma (LUSC).

TCGA-BRCA. (29) The TCGA-BRCA dataset includes 938 FFPE WSIs, of which 772 are diagnosed with Invasive Ductal Carcinoma (IDC) and 166 with Invasive Lobular Carcinoma (ILC).

TCGA-RCC. (29) The TCGA-RCC cohort contains 884 diagnostic WSIs covering three renal cell carcinoma subtypes: Chromophobe (TCGA-KICH), Clear Cell (TCGA-KIRC), and Papillary (TCGA-KIRP). The dataset includes 111 slides from 99 CRCC cases, 489 slides from 483 CCRCC cases, and 284 slides from 264 PRCC cases. On average, each slide contributed approximately 13,900 patches at $\times 20$ magnification.

PANDA. (30) The PANDA dataset consists of 12,625 prostate biopsy WSIs collected from six different institutions. The dataset includes 3,628 non-tissue/background slides, 3,151 non-epithelium/non-cancerous slides, 1,644 benign slides, and 4,202 cancerous slides. For our classification task, we focused on benign and cancerous slides to ensure a clinically meaningful evaluation.

3.2. Experimental Setup and Evaluation Metrics

All experiments were implemented in PyTorch and executed on a compute server equipped with four NVIDIA Tesla V100 GPUs and 32 CPU cores. Models were trained using a batch size of 256, the Adam optimizer with an initial learning rate of 0.001, and early stopping based on validation performance. We employed **5-fold cross-validation** for all datasets to ensure robust performance estimation. The token retention ratio r was used to tune the effective thresholds for both similarity-based merging and importance-guided pruning, with hyperparameters (merge utility weights, pruning ratios, and sparsity coefficient) optimized separately for each dataset. We report classification performance using Accuracy and Area Under the ROC Curve (AUC), where multi-class accuracy is computed as the average per-class accuracy and AUC is macro-averaged across classes.

3.3. Performance Comparison

We evaluated DTC-WSI on four benchmark WSI datasets—TCGA-NSCLC, TCGA-BRCA, TCGA-RCC, and PANDA—and compared it against leading MIL and token-efficient approaches. Table 1 reports the results. DTC-WSI achieves the highest accuracy on all datasets, reaching **98.3%** on TCGA-NSCLC, **97.4%** on TCGA-BRCA, **96.8%** on TCGA-RCC, and **94.8%** on PANDA. Across all benchmarks, it consistently outperforms strong baselines such as CLAM-MB, TransMIL, and HIPT by approximately **2–4%**, while using only **40%** of the original tokens—corresponding to a **5–10**$\times$ reduction in computational

Table 1: Comparison of **DTC-WSI** with prominent MIL and token-reduction approaches across four benchmark WSI datasets.

Model	**TCGA-NSCLC**		**TCGA-BRCA**		**TCGA-RCC**		**PANDA**	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
ABMIL (7)	94.7	95.4	93.4	94.1	92.6	93.5	91.2	92.1
CLAM-MB (8)	95.8	96.6	94.5	95.3	93.9	94.8	92.4	93.4
DSMIL (10)	95.3	96.1	94.0	94.8	93.4	94.2	91.9	92.9
TransMIL (9)	96.4	97.2	95.6	96.4	94.8	95.6	92.8	93.7
HIPT (11)	96.2	97.0	95.3	96.1	94.6	95.4	92.6	93.6
PANTHER (31)	96.8	97.6	96.0	96.8	95.2	96.1	93.3	94.2
SPT (32)	95.9	96.7	94.8	95.6	93.8	94.6	92.3	93.2
DTC-WSI (Ours)	**98.3**	**98.9**	**97.4**	**97.9**	**96.8**	**97.5**	**94.8**	**95.6**

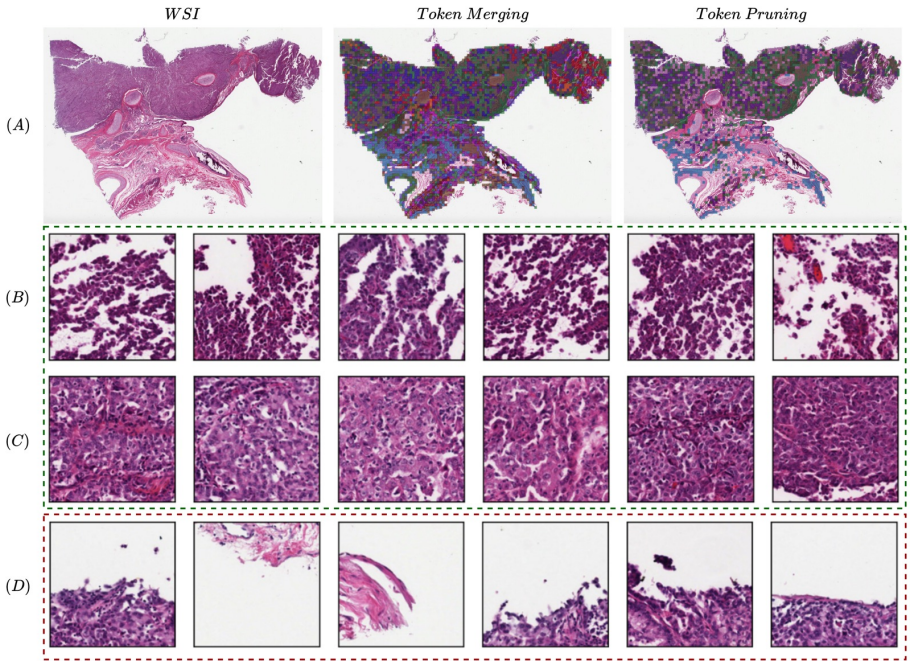

Figure 2: Visualization of multi-stage token compression in DTC-WSI: (A) Original WSI with post-merging and post-pruning heatmaps. (B–C) Similar patches merged into unified tokens (green), and (D) low-saliency patches removed by pruning (red).

load. We empirically found that a retention ratio of $r = 0.4$ provides the best balance between efficiency and accuracy (see Appendix B). These results demonstrate that dynamic multi-stage compression not only removes large-scale redundancy but also sharpens the model's discriminative focus by merging visually similar patches and pruning low-saliency regions under the guidance of the importance network.

Table 2: Effect of token compression on computational efficiency. DTC-WSI achieves substantial reductions in FLOPs, GPU memory, and inference time while improving predictive performance.

Tokens Retained	FLOPs (G)	GPU Memory (GB)	Inference Time (ms/WSI)	Speedup
r=0 (Baseline)	118.4	14.2	1280	1.0×
r=0.7	62.7	10.3	720	1.8×
r=0.5	38.4	8.1	420	3.0×
r=0.4 (Ours)	**24.3**	**6.4**	**240**	**5.3**×

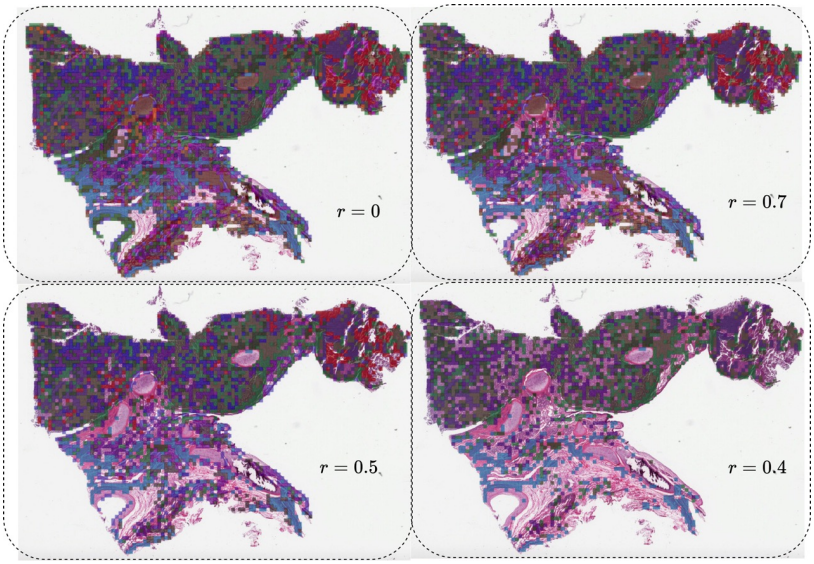

Figure 3: Visualization of multi-stage token compression in DTC-WSI across retention ratios $r \in \{0, 0.7, 0.5, 0.4\}$. Each panel shows the WSI after similarity-guided merging and importance-guided pruning, illustrating how token redundancy is reduced and salient regions are preserved as r decreases.

3.4. Computational Efficiency of Token Compression

In addition to improvements in predictive performance, DTC-WSI delivers substantial computational savings through its progressive token compression strategy. Table 2 summarizes how retaining fewer tokens reduces FLOPs, GPU memory, and inference time per WSI. With the full token set ($r = 0$), baseline inference requires 118.4 G FLOPs, 14.2 GB of memory, and 1280 ms per slide. Compressing to $r = 0.7$ nearly halves computation, yielding a 1.8× speedup. Further reduction to $r = 0.5$ continues to lower FLOPs and memory, reducing inference time to 420 ms. The full DTC-WSI configuration, retaining only $r = 0.4$ of the tokens, achieves the largest efficiency gains: FLOPs drop to 24.3 G, GPU memory falls to 6.4 GB, and inference time decreases to 240 ms—a 5.3× speedup over baseline. Notably, these improvements are achieved while *increasing* accuracy and AUC across all datasets. Overall, multi-stage token compression enhances representation quality while en-

abling fast, memory-efficient gigapixel WSI analysis, supporting scalable clinical deployment and real-time digital pathology workflows.

3.5. Ablation Studies

We conduct two ablation studies to evaluate the contributions of (1) **multi-stage token compression** and (2) the interaction between **similarity-guided merging** and **importance-guided pruning**. Results across all datasets show that each component of DTC-WSI is critical for achieving optimal performance.

Table 3 reports accuracy across retention ratios $r \in \{0, 0.7, 0.5, 0.4\}$. Performance consistently improves as redundant tokens are removed and the representation becomes more focused. For example, on *TCGA-NSCLC*, accuracy increases from **94.6%** (no compression) to **96.1%** (r=0.7), **97.1%** (r=0.5), and peaks at **98.3%** when retaining only 40% of tokens. Similar trends are observed on *TCGA-BRCA* (**93.9%** → **97.4%**), *TCGA-RCC* (**92.8%** → **96.8%**), and *PANDA* (**91.2%** → **94.8%**). These results confirm that multi-stage compression acts as a curriculum: early stages preserve global context, while later stages refine attention to diagnostically salient regions, improving both accuracy and robustness.

Table 4 further isolates the contributions of merging and pruning. Using **only merging** already yields substantial gains (e.g., **96.4%** on NSCLC), as redundant tissue regions are fused into compact representations. **Only pruning** likewise improves performance (e.g., **95.7%** on NSCLC) by removing low-saliency patches. However, replacing the importance network with **random token selection** substantially degrades accuracy across all datasets, demonstrating the need for learned saliency during compression.

The full DTC-WSI pipeline—combining similarity-guided merging, saliency-aware pruning, and a learned importance network—achieves the **highest accuracy on every dataset**, including **98.3%** (NSCLC), **97.4%** (BRCA), **96.8%** (RCC), and **94.8%** (PANDA). These ablations highlight that merging and pruning are complementary: merging eliminates redundancy, pruning removes noise, and importance-guidance ensures compression is both structured and diagnostically meaningful.

Table 3: Ablation study on multi-stage token compression across four datasets. Metrics reported as Accuracy (Acc) and AUC (%).

Compression Level	**TCGA-NSCLC**		**TCGA-BRCA**		**TCGA-RCC**		**PANDA**	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
No Compression (r=0)	94.6	96.1	93.9	95.0	92.8	94.1	91.2	92.5
Light Compression (r=0.7)	96.1	97.2	95.1	96.1	94.3	95.4	92.8	93.9
Moderate Compression (r=0.5)	97.1	98.0	96.2	97.0	95.4	96.4	93.8	94.9
Ours (r=0.4)	**98.3**	**98.9**	**97.4**	**97.9**	**96.8**	**97.5**	**94.8**	**95.6**

4. Visualization of Token Compression

Figure 2 illustrates the full multi-stage compression process performed by DTC-WSI. Panel (A) shows the original WSI along with overlaid heatmaps depicting the model output after similarity-guided token merging and after importance-guided pruning. Panels (B) and (C)

Table 4: Ablation study comparing merging and pruning strategies across four datasets. Metrics reported as Accuracy (Acc) and AUC (%).

Model Variant	**TCGA-NSCLC**		**TCGA-BRCA**		**TCGA-RCC**		**PANDA**	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
Only Merge	96.4	97.6	95.3	96.3	94.5	95.7	93.0	94.2
Only Prune	95.7	96.9	94.7	95.8	93.9	95.1	92.4	93.5
Random Token Selection	94.2	95.4	93.1	94.3	92.5	93.8	90.9	92.1
Ours (Merge + Prune)	**98.3**	**98.9**	**97.4**	**97.9**	**96.8**	**97.5**	**94.8**	**95.6**

present examples of visually similar patches that are merged into unified representations; these merged groups are highlighted with green borders, demonstrating how redundant regions—such as uniform stromal areas or repeated tumor patterns—are effectively consolidated. Panel (D) displays patches removed through importance-guided pruning, marked with red borders, revealing low-saliency regions that contribute minimally to the slide-level prediction. Overall, these visualizations show that DTC-WSI performs structured, interpretable compression: reducing redundancy through merging while selectively pruning non-informative regions, ultimately preserving the most diagnostically meaningful tissue patterns.

Figure 3 illustrates how DTC-WSI progressively compresses a whole-slide image as the token retention ratio is reduced from $r = 0$ (no compression) to $r = 0.7$, $r = 0.5$, and $r = 0.4$. For each retention level, we visualize the WSI after similarity-guided merging and importance-guided pruning. At $r = 0$, all extracted patches are preserved, resulting in a dense and highly redundant representation. As r decreases, visually homogeneous regions (e.g., stroma, fat, repeated tumor textures) are merged into unified tokens, while low-saliency patches are pruned, yielding a more compact and interpretable representation. By $r = 0.4$, the model retains only the most discriminative tissue regions, producing a sparse yet diagnostically meaningful token set. These visualizations demonstrate how multi-stage compression removes redundancy while concentrating model capacity on morphologically informative regions, enabling both efficiency and performance gains.

5. Conclusion

We presented **DTC-WSI**, a scalable framework for token-efficient whole-slide image analysis. By combining similarity-guided merging with importance-guided pruning in a progressive multi-stage pipeline, DTC-WSI removes redundancy while preserving diagnostically essential information. The method supports differentiable compression during training and deterministic reduction at inference, achieving **5–10× token reduction**, **5.3× faster inference**, and **40% lower memory usage** without sacrificing accuracy. Across four benchmark datasets, DTC-WSI improves classification performance by **2–4%**, demonstrating that compression can enhance representation quality rather than degrade it.

More broadly, our results show that *structured token merging*, guided by learned importance, offers a powerful alternative to pruning alone for large-scale vision tasks. DTC-WSI opens new directions for efficient transformer design, scalable pathology workflows, and adaptive token allocation in high-resolution visual reasoning.

References

[1] El Nahhas, Omar SM, Marko van Treeck, Georg Wölflein, Michaela Unger, Marta Ligero, Tim Lenz, Sophia J. Wagner et al. "From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology." Nature Protocols 20, no. 1 (2025): 293-316.

[2] Srinidhi, Chetan L., Ozan Ciga, and Anne L. Martel. "Deep neural network models for computational histopathology: A survey." Medical image analysis 67 (2021): 101813.

[3] Cui, M., Zhang, D.Y.: Artificial intelligence and computational pathology. Lab. Invest. 101, 412–422 (2021)

[4] El Nahhas, O. S., van Treeck, M., Wölflein, G., Unger, M., Ligero, M., Lenz, T., ... Kather, J. N. (2025). From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. Nature Protocols, 20(1), 293-316.

[5] Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B.: Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. 2, 147–171 (2009)

[6] Bruny´e, T.T., Mercan, E., Weaver, D.L., Elmore, J.G.: Accuracy is in the eyes of the pathologist: the visual interpretive process and diagnostic accuracy with digital whole slide images. J. Biomed. Info. 66, 171–179 (2010)

[7] Ilse, M., Tomczak, J.,Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)

[8] Lu, M.Y., Williamson, Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. 5(6), 555–570 (2021)

[9] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: transformer based correlated multiple instance learning for whole slide image classification. Adv. Neural Inf. Process. Syst. 34, 2136–2147 (2021)

[10] Li, Bin, Yin Li, and Kevin W. Eliceiri. "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.

[11] Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G. and Mahmood, F., 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16144-16155).

[12] Rahman, Tawsifur, Alexander S. Baras, and Rama Chellappa. "Evaluation of a Task-Specific Self-Supervised Learning Framework in Digital Pathology Relative to Transfer Learning Approaches and Existing Foundation Models." Modern Pathology 38, no. 1 (2025): 100636.

[13] Yang, X., et al.: Virtual stain transfer in histology via cascaded deep neural networks. ACS Photonics 9(9), 3134–3143 (2022)

[14] Kapse, Saarthak, Pushpak Pati, Srijan Das, Jingwei Zhang, Chao Chen, Maria Vakalopoulou, Joel Saltz, Dimitris Samaras, Rajarsi R. Gupta, and Prateek Prasanna. "SI-MIL: Taming Deep MIL for Self-Interpretability in Gigapixel Histopathology." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11226-11237. 2024.

[15] Tang, Wenhao, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. "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.

[16] T. Rahman, A. S. Baras and R. Chellappa, "CEMIL: Contextual Attention Based Efficient Weakly Supervised Approach for Histopathology Image Classification," 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 2025, pp. 4248-4257

[17] Yue, Longzhe, Haixing Li, and Wanying Huang. "Hierarchical Cross-Scale Attention-Based Multi-Instance Learning for Whole Slide Image Classification." In 2025 8th International Conference on Computer Information Science and Application Technology (CISAT), pp. 493-497. IEEE, 2025.

[18] Sharma, Y., Shrivastava, A., Ehsan, L., Moskaluk, C.A., Syed, S. and Brown, D., 2021, August. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In Medical imaging with deep learning (pp. 682-698). PMLR.

[19] Tang, Q., Zhang, B., Liu, J., Liu, F. and Liu, Y., 2023. Dynamic token pruning in plain vision transformers for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 777-786).

[20] Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J. and Hsieh, C.J., 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in neural information processing systems, 34, pp.13937-13949.

[21] Jiang, P., Li, X., Shen, H., Chen, Y., Wang, L., Chen, H., Feng, J. and Liu, J., 2023. A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis. Artificial Intelligence Review, 56(Suppl 2), pp.2687-2758.

[22] Shi, J.Y., Wang, X., Ding, G.Y., Dong, Z., Han, J., Guan, Z., Ma, L.J., Zheng, Y., Zhang, L., Yu, G.Z. and Wang, X.Y., 2021. Exploring prognostic indicators in the pathological images of hepatocellular carcinoma based on deep learning. Gut, 70(5), pp.951-961.

[23] Lyu, W., Hu, Q., Qi, K., Shi, Z., Huang, W., Gupta, S. and Chen, C., 2025. Efficient whole slide pathology vqa via token compression. arXiv preprint arXiv:2507.14497.

[24] Bolya, D. and Hoffman, J., 2023. Token merging for fast stable diffusion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4599-4603).

[25] Zhen, T., Cao, J., Sun, X., Pan, J., Ji, Z. and Pang, Y., 2025. Token-aware and step-aware acceleration for Stable Diffusion. Pattern Recognition, 164, p.111479.

[26] Hu, T., Li, L., van de Weijer, J., Gao, H., Shahbaz Khan, F., Yang, J., Cheng, M.M., Wang, K. and Wang, Y., 2024. Token merging for training-free semantic binding in text-to-image synthesis. Advances in Neural Information Processing Systems, 37, pp.137646-137672.

[27] Wu, H., Xu, J., Le, H. and Samaras, D., 2025. Importance-based token merging for efficient image and video generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4983-4995).

[28] Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G. and Parwani, A.V., 2024. A visual-language foundation model for computational pathology. Nature medicine, 30(3), pp.863-874.

[29] Tomczak, K., Czerwińska, P. and Wiznerowicz, M., 2015. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia, 2015(1), pp.68-77. (2015)

[30] Bulten, W., Kartasalo, H., Vink, R. and Hulsbergen-van de Kaa, C. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nature medicine, 28(1), pp.154-163. (2022)

[31] Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G. and Mahmood, F., 2024. 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).

[32] Hou, Xinhai, Cheng Jiang, Akhil Kondepudi, Yiwei Lyu, Asadur Zaman Chowdury, Honglak Lee, and Todd C. Hollon. "A self-supervised framework for learning whole slide representations." arXiv preprint arXiv:2402.06188 (2024)

Appendix A.

Whole Slide Image Preprocessing

Whole slide image (WSI) preprocessing begins with automated tissue segmentation. Each WSI is first loaded into memory at a downsampled resolution, such as 20×, and converted from RGB to HSV colorspace. Tissue regions (foreground) are identified by thresholding the saturation channel after applying median blurring to smooth edges. A binary mask is then generated and refined using morphological closing to eliminate small gaps and holes. The contours of detected tissue regions are filtered based on an area threshold, ensuring only relevant regions are retained for further processing. The segmentation mask for each

slide is also available for optional visual inspection. To facilitate manual adjustments, a human-readable text file is generated, listing processed files along with editable segmentation parameters. Once segmentation is complete, 256×256 patches are extracted from within the segmented contours at the specified magnification. These patches, along with their coordinates and slide metadata, are stored in the HDF5 hierarchical data format. The number of extracted patches per slide varies significantly—ranging from hundreds in biopsy slides at 20× magnification to hundreds of thousands in large resection slides at 40× magnification.

Appendix B. Ablation study

Comparison of Different Threshold Values

The extended ablation in Table 5 evaluates DTC-WSI under a wide range of token retention ratios ($r \in [0.3, 0.8]$) across four benchmark datasets. Performance improves consistently as redundant tokens are removed, with accuracy rising steadily from $r = 0.8$ to $r = 0.5$ on all cohorts. The model achieves its best results at $r = 0.4$, reaching **98.3%** (NSCLC), **97.4%** (BRCA), **96.8%** (RCC), and **94.8%** (PANDA), demonstrating that moderate compression enhances discriminative focus while preserving essential morphology. When compression becomes too aggressive ($r = 0.3$), performance drops sharply—e.g., NSCLC declines from **98.3%** to **90.4%**—indicating loss of critical diagnostic tokens. These results highlight a clear U-shaped trend: light compression reduces redundancy, moderate compression maximizes accuracy, and over-compression degrades performance. Overall, the study confirms that DTC-WSI benefits most from token retention around $r = 0.4$, where efficiency and predictive power are jointly optimized.

Table 5: Extended ablation study evaluating token retention ratios across four datasets. Metrics reported as Accuracy (Acc) and AUC (%).

Retention Ratio (r)	**TCGA-NSCLC**		**TCGA-BRCA**		**TCGA-RCC**		**PANDA**	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
r = 0.8	95.6	96.6	94.6	95.6	93.6	94.7	92.0	93.3
r = 0.7	96.1	97.2	95.1	96.1	94.3	95.4	92.8	93.9
r = 0.6	96.6	97.5	95.6	96.6	95.0	95.9	93.3	94.4
r = 0.5	97.1	98.0	96.2	97.0	95.4	96.4	93.8	94.9
r = 0.4 (Best)	**98.3**	**98.9**	**97.4**	**97.9**	**96.8**	**97.5**	**94.8**	**95.6**
r = 0.3	90.4	92.2	93.5	94.2	88.8	89.8	85.9	86.8

Appendix C. Visualization of Token Compression

We provide visualizations to illustrate how DTC-WSI compresses WSIs while preserving diagnostically important tissue. In Figure. 2, which shows the original WSI of lung adenocarcinoma, the second panel visualizes similarity-guided merging by assigning identical interior and boundary colors to patches that are merged into a single token. This reveals how homogeneous tissue regions—such as smooth stroma or repeated tumor textures—are

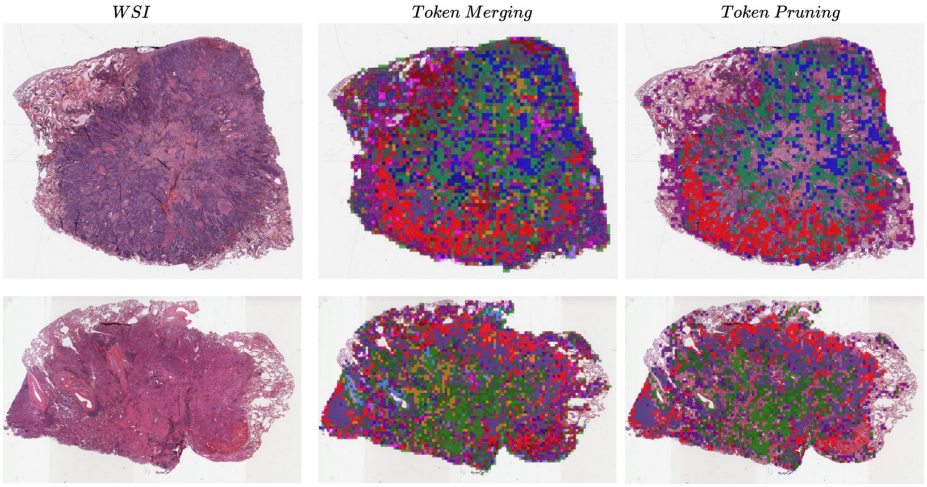

Figure 4: Visualization of the multi-stage token compression in DTC-WSI. (Left) Original WSI thumbnail. (Middle) Similarity-guided merging groups redundant patches into shared representations(Patches with the same inner and border color are merged together.) (Right) Importance-guided pruning removes low-saliency tokens.

consolidated into compact groups, while heterogeneous or diagnostically subtle regions remain unmerged. The third panel displays the result of importance-guided pruning, where tokens with low saliency scores are removed entirely, leaving a focused set of highly informative patches concentrated around tumor-rich or otherwise relevant regions. Together, these visualizations demonstrate that DTC-WSI performs structured and interpretable compression, reducing redundancy while retaining the critical morphological patterns needed for accurate WSI classification.

Appendix D. Algorithm of DTC-WSI

Algorithm 1: Dynamic Token Compression for Whole-Slide Images (DTC-WSI)

Input: Patch features $H^{(0)} = \{h_i^{(0)}\}_{i=1}^{N^{(0)}}$, # stages T, target token counts $\{N^{(t)}\}_{t=1}^{T}$, mode $\in \{\text{train}, \text{infer}\}$

Output: Compressed token set $H^{(T)}$

for $t = 0$ **to** $T-1$ **do**

  $N^{(t)} \leftarrow |H^{(t)}|$; `// current #tokens`

  `/* 1. Importance estimation */`

  for $i = 1$ **to** $N^{(t)}$ **do**

    $s_i^{(t)} \leftarrow g_\phi(h_i^{(t)})$; `// importance score`

  end

  $\alpha^{(t)} \leftarrow \mathrm{softmax}(s^{(t)})$; `// normalized importance`

  `/* 2. Bipartite soft matching for token fusion */`

  `// Interleaved partition: odd indices` $\rightarrow A$, `even indices` $\rightarrow B$

  $A \leftarrow [1, 3, 5, \dots]$, $\quad B \leftarrow [2, 4, 6, \dots]$ Let $L = \min(|A|, |B|)$

  for $k = 1$ **to** L **do**

    $i \leftarrow A_k, \quad j \leftarrow B_k \;\; \mathrm{sim}_{ij} \leftarrow \dfrac{\langle h_i^{(t)}, h_j^{(t)} \rangle}{\|h_i^{(t)}\| \, \|h_j^{(t)}\|} \;\; u_{ij}^{(t)} \leftarrow \lambda \, \mathrm{sim}_{ij} - (1-\lambda)\,|\alpha_i^{(t)} - \alpha_j^{(t)}|$

  end

  `// Number of pairs to merge`

  $K^{(t)} \leftarrow \max\left(0,\, N^{(t)} - N^{(t+1)}\right)$ Select top-$N^{(t)}$ pairs $\mathcal{P}^{(t)}$ sorted by $u_{ij}^{(t)}$

  `/* 3. Merge selected pairs */`

  Initialize $H_{\text{merge}}^{(t+1)} \leftarrow \emptyset$, mark all indices as "unassigned"

  foreach $(i, j) \in \mathcal{P}^{(t)}$ *with both* i, j *unassigned* **do**

    $\tilde{h}_i^{(t)} \leftarrow \dfrac{\alpha_i^{(t)} h_i^{(t)} + \alpha_j^{(t)} h_j^{(t)}}{\alpha_i^{(t)} + \alpha_j^{(t)}}$ Add $\tilde{h}_i^{(t)}$ to $H_{\text{merge}}^{(t+1)}$ Mark i and j as "assigned"

  end

  `/* 4. Carry over unmerged tokens */`

  $H_{\text{carry}}^{(t+1)} \leftarrow \{h_k^{(t)} \mid k \text{ unassigned}\}$ $H_{\text{raw}}^{(t+1)} \leftarrow H_{\text{merge}}^{(t+1)} \cup H_{\text{carry}}^{(t+1)}$

  `/* 5. Importance-guided pruning */`

  if *mode = train* **then**

    `// soft, differentiable pruning`

    foreach $h_k^{(t+1)} \in H_{raw}^{(t+1)}$ **do**

      $m_k^{(t)} \leftarrow \sigma\big(\gamma(\alpha_k^{(t)} - \tau)\big)$ $h_k^{(t+1)} \leftarrow m_k^{(t)}\, h_k^{(t+1)}$

    end

    $H^{(t+1)} \leftarrow H_{\text{raw}}^{(t+1)}$

  end

  else

    `// hard top-`$N^{(t+1)}$ `pruning at inference`

    Rank all $h_k^{(t+1)} \in H_{\text{raw}}^{(t+1)}$ by $\alpha_k^{(t)}$ $H^{(t+1)} \leftarrow \mathrm{TopK}(H_{\text{raw}}^{(t+1)}, \alpha^{(t)}, N^{(t+1)})$

  end

end

return $H^{(T)}$