Context Matters: Query-aware Dynamic Long Sequence Modeling of Gigapixel Images

Zhengrui Guo [1 2] **Qichen Sun** [3] **Jiabo Ma** [1] **Lishuang Feng** [2] **Jinzhuo Wang** [3] **Hao Chen** [1]

Abstract

Whole slide image (WSI) analysis presents significant computational challenges due to the massive number of patches in gigapixel images. While transformer architectures excel at modeling long-range correlations through self-attention, their quadratic computational complexity makes them impractical for computational pathology applications. Existing solutions like local-global or linear self-attention reduce computational costs but compromise the strong modeling capabilities of full self-attention. In this work, we propose **Querent**, *i.e.*, the **quer**y-awar**e** long co**nt**extual dynamic modeling framework, which achieves a theoretically bounded approximation of full self-attention while delivering practical efficiency. Our method adaptively predicts which surrounding regions are most relevant for each patch, enabling focused yet unrestricted attention computation only with potentially important contexts. By using efficient region-wise metadata computation and importance estimation, our approach dramatically reduces computational overhead while preserving global perception to model fine-grained patch correlations. Through comprehensive experiments on biomarker prediction, gene mutation prediction, cancer subtyping, and survival analysis across over 10 WSI datasets, our method demonstrates superior performance compared to the state-of-the-art approaches. Codes are here.

1. Introduction

Computational pathology (CPath) represents a transformative shift in clinical diagnostics, leveraging artificial intelligence and deep learning to analyze the growing collections of whole slide images (WSIs) from medical facilities (Van der Laak et al., 2021; Cui & Zhang, 2021; Zheng et al., 2025). By digitalizing traditional pathology workflows, this emerging field enhances clinical decision-making through more standardized diagnoses, enables the identification of new biomarkers, and helps predict treatment outcomes for patients (Niazi et al., 2019; Song et al., 2023). These WSIs, also known as gigapixel images, typically contain between $10,000^2 \sim 100,000^2$ pixels. The challenge lies in identifying critical diagnostic features that may be dispersed across various tissue regions within these highly informative images, *i.e.*, analogous to finding a needle in a haystack (Jin et al., 2023). These unique challenges have driven the development of multi-instance learning (MIL), a weakly-supervised learning paradigm for WSI analysis (Campanella et al., 2019; Lu et al., 2021; Shao et al., 2021; Xu & Chen, 2023; Zhou & Chen, 2023; Yang et al., 2024).

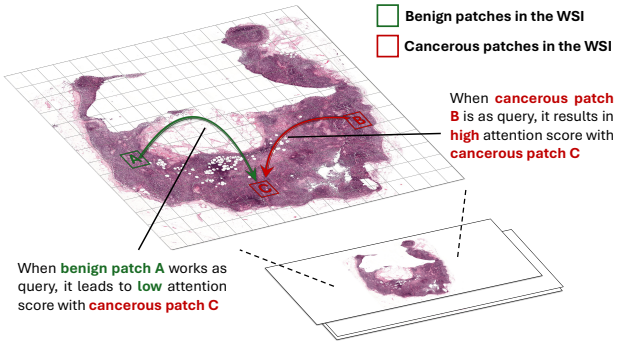

Figure 1. Illustration of context-dependent patch relationships in whole slide images. When a benign patch (A) interacts with cancerous patch C, it shows low correlation, while a cancerous patch (B) shows high correlation with patch C. This demonstrates how the same patch (C) can have fundamentally different relationships with other patches depending on the biological context.

MIL performs slide-level analysis through three key steps: (1) segmenting and cropping tissues in a WSI into smaller image patches; (2) extracting representation features from these patches using a pretrained encoder (typically a CPath foundation model (Chen et al., 2024; Lu et al., 2024; Ma et al., 2024; Xu et al., 2024; Du et al., 2025; Xiong et al., 2025a)); and (3) aggregating the patch features through

[1]Hong Kong University of Science and Technology, HK, China [2]Beijing Institute of Collaborative Innovation, Beijing, China [3]Peking University, Beijing, China. Correspondence to: Zhengrui Guo <zguobc@connect.ust.hk>, Hao Chen <jhc@cse.ust.hk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

specific principles to obtain the slide-level representation. Among various feature aggregation approaches, transformer architectures have emerged as a particularly promising solution due to their powerful ability to model long-range dependencies through self-attention mechanisms (Vaswani, 2017; Dosovitskiy, 2020). This capability is especially valuable in WSI analysis, where diagnostically relevant features often manifest through complex spatial relationships between distant tissue regions. However, the quadratic computational complexity $\mathbf{O}(n^2)$ of the standard transformer's self-attention mechanism presents a significant challenge when applied to WSIs, as a typical slide can contain thousands to tens of thousands of patches (Wang et al., 2023; Song et al., 2024b; Guo et al., 2025; Xiong et al., 2025b).

To address this computational barrier, several adaptations of transformer architectures have been proposed for WSI analysis. These include linear attention methods that reduce complexity to $\mathbf{O}(n)$ (Shao et al., 2021), and local-global attention approaches that restrict attention computation to predetermined spatial patterns (Chen et al., 2022; Li et al., 2024a). While these modifications successfully reduce computational overhead, they inevitably compromise the transformer's inherent modeling capabilities. As demonstrated by Dao et al. (2022) and Han et al. (2023), linear approximation of self-attention leads to only sub-optimal modeling performance. Meanwhile, local-global attention makes strong assumptions about which spatial relationships are important, failing to adapt to the highly variable and context-dependent nature of pathological features in WSIs.

These limitations motivate the need for a more adaptive approach to modeling inter-patch relationships in WSI analysis. A key observation is that the relevance of surrounding tissue regions varies significantly depending on which specific region is being examined (Heindl et al., 2015; Yuan, 2016). As shown in Fig. 1, when analyzing a tumor boundary region, nearby regions showing the tumor-stroma interface might be highly relevant, while distant regions of normal tissue might be less informative. Conversely, when examining an area of inflammation, regions with similar inflammatory patterns across the slide might be more relevant than adjacent but histologically different regions. This context-dependent nature of patch relationships suggests that an ideal attention mechanism should dynamically prioritize relevant interactions for each query patch while maintaining the capability to model long-range dependencies when necessary.

Based on this insight, we propose **Querent**, a novel framework for dynamic long-range contextual modeling of gigapixel images through adaptive determination of patch relationships. Our approach maintains the modeling power of full attention while achieving computational efficiency through dynamic sparsification. Rather than using fixed patterns or uniform approximations, it estimates the potential importance of patch relationships through efficient region metadata computation and selectively applies full attention to the most relevant interactions. This query-aware strategy allows each patch to have its unique attention pattern, better capturing the heterogeneous nature of histological features while remaining computationally tractable for gigapixel WSIs. Theoretically, we prove that Querent's query-aware attention mechanism maintains expressiveness within a small constant bound of full self-attention. Empirically, we demonstrate Querent's effectiveness across multiple CPath tasks, including biomarker prediction, gene mutation prediction, cancer subtyping, and survival prediction, where it consistently outperforms state-of-the-art models on over 10 WSI datasets.

The main contributions of this work are as follows: (1) We propose a novel query-aware attention mechanism that dynamically adapts to each patch's unique context in gigapixel WSIs, maintaining full attention's expressiveness while achieving computational efficiency; (2) We develop efficient region-level metadata summarization and importance estimation modules that enables dynamic sparsification of attention patterns while preserving modeling capabilities; (3) We establish Querent's effectiveness through theoretical bounds on its expressiveness and extensive empirical validation across diverse CPath tasks.

2. Related Work

Given the massive image size of WSIs and GPU memory restrictions (Araujo et al., 2019), researchers have widely adopted MIL for WSI analysis. Recent MIL-based approaches have achieved promising results in diagnosing diseases and predicting patient outcomes under the formulation of weakly-supervised learning (Campanella et al., 2019; Chikontwe et al., 2020; Li et al., 2021b; Xiang et al., 2022; Hou et al., 2022; Zheng et al., 2022; Wang et al., 2022; Yu et al., 2023; Xiong et al., 2023; Lin et al., 2023; Xiong et al., 2024a; Song et al., 2024a; Xiong et al., 2024b).

MIL primarily addresses the challenge of aggregating patch-level features into a comprehensive slide-level representation for diagnostic purposes. Early MIL approaches employed straightforward, non-parametric aggregation techniques, such as max and mean pooling operations, to combine these features (Campanella et al., 2019). Recent advances in MIL have emphasized the development of more sophisticated aggregation strategies to effectively capture diagnostic patterns scattered across numerous patches. ABMIL (Ilse et al., 2018) introduces an attention-based framework that computes importance weights for each patch using a side network during the aggregation process. CLAM (Lu et al., 2021) enhanced this attention mechanism by incorporating clustering-constrained learning and flexible attention branches to enhance WSI classification performance.

DSMIL (Li et al., 2021a) develops a dual-stream architecture that capitalizes on the hierarchical structure of WSIs by integrating features across multiple magnification levels. Further, DTFD-MIL (Zhang et al., 2022) introduces a novel double-tier framework that leverages pseudo-bags to maximize feature utilization and minimize the bag-instance imbalance problem in WSI analysis. Taking a graph-based perspective, WiKG (Li et al., 2024b) reformulates WSI analysis by representing patches as nodes in a knowledge graph and utilizing head-to-tail embeddings to generate dynamic graph representations. Combined with the State Space Models (Gu & Dao, 2023), MambaMIL (Yang et al., 2024) features a Sequence Reordering Mamba module that processes instance sequences through both original and reordered pathways to enhance long-range dependency modeling while maintaining linear complexity. The field has further evolved with the integration of multimodal information, as researchers have begun incorporating pathology image captions (Lu et al., 2023), diagnostic reports (Guo et al., 2024), as well as genomic profiles (Sun et al., 2025) to enhance both interpretability and diagnostic accuracy.

Building upon these advancements, another significant line of research has focused on leveraging Transformer architectures to model long-range dependencies among patches in WSIs. Despite the effectiveness of self-attention in modeling long-range correlation, the quadratic computational complexity of standard Transformer architectures poses significant challenges when processing tens of thousands of patches in a WSI. To address this limitation, Transformer-based MIL methods could be categorized into two classes, *i.e.*, linear approximation (Shao et al., 2021) and local-global attention (Chen et al., 2022; Guo et al., 2024; Li et al., 2024a). TransMIL (Shao et al., 2021), a pioneering work in linear approximation, adopts the Nyströmformer (Xiong et al., 2021) to achieve linear complexity in WSI modeling. However, this approximation strategy often leads to sub-optimal performance due to its inherent limitations in capturing pairwise token interactions, creating an information bottleneck that compromises the model's ability to learn complex dependencies (Dao et al., 2022). In contrast, local-global attention methods like HIPT (Chen et al., 2022) introduce a hierarchical approach, using a three-level Transformer architecture to progressively aggregate information from cellular features to tissue phenotypes through non-overlapping region slicing. Similarly, HistGen (Guo et al., 2024), RRT-MIL (Tang et al., 2024), and LongMIL (Li et al., 2024a) employ a two-stage strategy, first processing patches within local attention windows before pooling them to enable global attention across the reduced sequence. However, these methods rely on fixed window sizes, which fail to capture the inherently adaptive nature of pathological analysis: malignant regions often require attention to distant but visually similar areas, while normal tissue typically benefits from focusing on local contextual patterns.

This observation motivates our query-aware dynamic long sequence modeling approach, which adaptively determines relevant regions for each patch based on its pathological characteristics rather than using predetermined fixed attention patterns. Our method efficiently computes region-level metadata through min-max networks to estimate the importance of different tissue regions for each (query) patch. By focusing only on the most relevant regions for each patch, we maintain the expressiveness of full attention while significantly reducing computational cost.

3. Methods

3.1. Problem Formulation and Solution Overview

Problem Formulation: Given a WSI W, we first partition it into a set of non-overlapping smaller patches $\{p_1, p_2, ..., p_N\}$ at a fixed magnification level. Each patch is then processed through a pre-trained encoder $\phi(\cdot)$ to obtain feature representations. Specifically, a CPath foundation model named PLIP (Huang et al., 2023) is used for this purpose. Formally, for each patch p_i, we obtain its feature representation $x_i = \phi(p_i)$, where $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector. This results in a bag of instances $X = \{x_1, x_2, ..., x_N\}$ representing the entire WSI. The goal of WSI analysis is to learn a mapping function $f : \mathcal{X} \to \mathcal{Y}$ that predicts the slide-level label y based on all patches within the WSI through a weakly-supervised manner. The key challenge lies in effectively aggregating information from thousands of patches while capturing both their specific contexts and inter-patch semantic relationships.

Solution Overview: Here, we describe the overview of our method comprising 4 major steps as depicted in Fig. 2.

Step 1: For each WSI, we partition the WSI into local regions, where each region contains multiple patches. For each region, we compute region-level metadata through a summarization mechanism. This metadata encapsulates the key characteristics of each region through statistical measures (*e.g.*, min/max features) and serves as a compact representation for importance estimation.

Step 2: Given a query patch, we leverage the pre-computed region-level metadata to identify the top-K most relevant regions through importance scoring efficiently. This enables prioritized processing of regions most likely to contain meaningful contextual information for the current query.

Step 3: Then, we employ query-aware selective attention to process the query with the most relevant regions. This step involves computing dense self-attention between the query and patches within selected regions, allowing for detailed analysis of patch-to-patch relationships while maintaining computational efficiency. The selection is guided by the

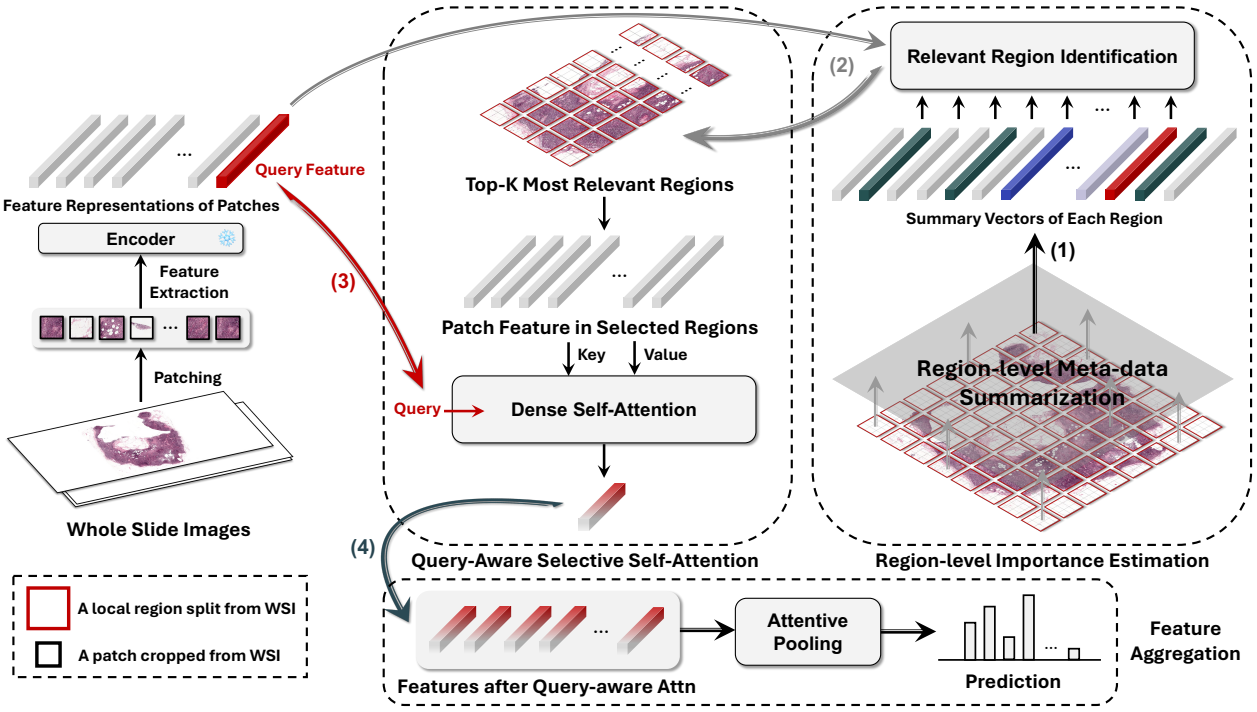

Figure 2. Illustration of the proposed **Querent** framework, which models a WSI via four key steps: (1) region-level metadata summarization from the partitioned WSI, detailed in Fig. 3, (2) identification of relevant regions for query patches through efficient importance scoring, (3) query-aware selective self-attention computation between query patch and patches in selected regions, and (4) feature aggregation with attentive pooling for final prediction. The framework enables dynamic modeling of long-range contextual relationships in gigapixel WSIs through efficient region relevance identification and query-aware selective attention computation.

relevance scores computed using the region-level metadata.

Step 4: Finally, we aggregate the refined features through an attentive pooling mechanism to generate slide-level predictions, which combines the contextually enhanced patch representations while emphasizing the most diagnostically relevant features.

3.2. Querent for dynamic long sequence WSI modeling

3.2.1. Region-level Metadata Summarization

Given the extracted patch features $X = \{x_1, x_2, ..., x_N\}$, we first organize them into regions $\mathcal{R} = \{R_1, R_2, ..., R_M\}$, where each region R_i contains a fixed number of K patches.

For each region R_i, we compute metadata vectors that capture the statistical characteristics of all patches within that region, as illustrated in Fig. 3. Specifically, we compute two types of summary vectors: minimum and maximum feature values across all patches in the region. Formally, for region R_i containing patches $\{x_{i1}, x_{i2}, ..., x_{iK}\}$, we compute:

$$m_i^{\min} = \min_{j \in \{1...K\}} x_{ij}, \quad m_i^{\max} = \max_{j \in \{1...K\}} x_{ij} \tag{1}$$

where $m_i^{\min}, m_i^{\max} \in \mathbb{R}^d$ represent the element-wise minimum and maximum values across all patches in region R_i. These summary vectors are then transformed through learnable projections:

$$\hat{m}_i^{\min} = f_{\min}(m_i^{\min}), \quad \hat{m}_i^{\max} = f_{\max}(m_i^{\max}) \tag{2}$$

where $f_{\min}$ and $f_{\max}$ are neural networks that project the summary vectors into a shared embedding space. This projection allows the metadata to be directly comparable with query features in the subsequent importance estimation step. The resulting region-level metadata provides a compact yet informative representation of each region's content, enabling efficient relevance assessment without requiring exhaustive patch-level computations.

3.2.2. Region Importance Estimation for Query

Given a query patch feature $q \in \mathbb{R}^d$, our goal is to efficiently identify the most relevant regions for this query using the pre-computed region-level metadata. For each region R_i, we estimate its importance score based on the potential interaction between the query and its metadata vectors.

Specifically, we first project the query feature into the same embedding space as the region metadata $\hat{q} = f_q(q)$, where f_q is a learnable projection network. The importance score for region R_i is then computed by evaluating the maximum possible interaction between the query and the region's meta-

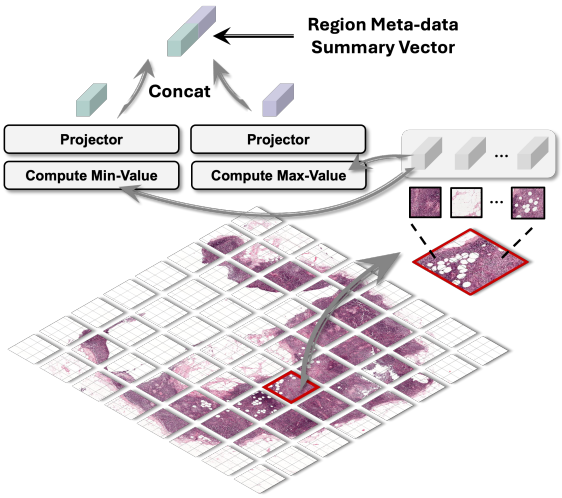

Figure 3. Illustration of the region-level metadata summarization process. Each region from the WSI is represented by summary vectors computed from its constituent patches. These summary vectors capture the statistical characteristics (minimum and maximum values) across all patches within each region, providing an efficient representation for subsequent importance estimation.

data bounds:

$$s_i = \max(|\langle \hat{q}, \hat{m}_i^{\min} \rangle|, |\langle \hat{q}, \hat{m}_i^{\max} \rangle|) \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product operation. The importance score s_i provides an upper bound on the potential relevance of any patch within region R_i to the current query. Based on these scores, we select the top-K regions with the highest importance scores:

$$\mathcal{R}_q = \text{TopK}(\{(R_i, s_i)\}_{i=1}^{M}) \tag{4}$$

This efficient scoring mechanism allows us to identify the most promising regions for detailed attention computation without examining every patch, significantly reducing the computational complexity while maintaining the ability to capture long-range dependencies.

3.2.3. Query-Aware Selective Attention

After identifying the relevant regions $\mathcal{R}q$ for a query patch q, we compute dense self-attention between the query and patches within these selected regions. For a selected region $R_i \in \mathcal{R}q$, we first compute query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) representations through linear projections, followed by attention score computation:

$$[\mathbf{Q}, \mathbf{K}, \mathbf{V}] = \mathbf{W}_{\text{qkv}}(x), \quad \mathbf{A} = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_h}}) \tag{5}$$

where $\mathbf{W}_{\text{qkv}} \in \mathbb{R}^{d \times 3d}$ is a learnable parameter matrix, x represents the concatenation of the query patch and patches from the selected region, and d_h is the dimension of each attention head. The output of the attention layer and multi-head attention are computed as:

$$\mathbf{O} = \mathbf{A}\mathbf{V} \tag{6}$$

$$\mathbf{O}_h = \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\mathbf{O}_1, \ldots, \mathbf{O}_H)\mathbf{W}_O \tag{7}$$

where $\mathbf{W}_O \in \mathbb{R}^{Hd_h \times d}$ is a learnable projection matrix. Through this selective attention mechanism, each query patch attends only to patches within the most relevant regions, achieving a balance between computational efficiency and comprehensive contextual modeling. The computational complexity is reduced from $O(N^2)$ to $O(NK)$, where N is the total number of patches and K is the number of patches in selected regions.

The selective attention mechanism described above provides an efficient approximation of full self-attention while maintaining its key properties. To formally characterize the relationship between our query-aware selective attention and standard self-attention, we establish the following theoretical guarantee. This theorem demonstrates that under reasonable conditions regarding the input distribution and model parameters, our selective attention mechanism can approximate full self-attention with bounded error. Specifically, we show:

Theorem 3.1 (Query-Aware Attention Approximation)**.** *Let* $\mathbf{A}$ *be the query-aware attention matrix (Def. B.2), and* $\mathbf{B}$ *be the full self-attention matrix (Def. B.1). Assume attention scores decay exponentially with spatial distance:* $\exp(-\alpha d(i,j))$ *bounds the attention score decay for distance* $d(i,j)$*. For any input sequence* $\mathbf{X}$ *with* L *patches and* d *dimensions, there exist* ***random projection matrices*** $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d \times d}$ *such that* $\|\mathbf{A} - \mathbf{B}\|_F \leq \left(2 + \frac{B}{\sqrt{d}}\right)\epsilon$ *with probability at least* $1 - \delta$*, provided:*

1. *The hidden dimension satisfies, and the number of selected regions per query satisfies:*

$$d \geq C_1 \cdot \frac{\log(L/\delta)}{\epsilon^2}, \quad k \geq \frac{C_2}{\alpha} \cdot \log\left(\frac{1}{\epsilon}\right)$$

   where $C_1 = 8B^4$ *(from JL inner-product preservation (Kaban, 2015)), and* $C_2 = 2$*.*

2. *For each region* R_i*, the diameter satisfies:*

$$\text{diam}(R_i) \leq \min\left(\frac{\epsilon}{L \cdot \sqrt{d}}, \frac{1}{\alpha}\right)$$

3. *Regions are spatially separated such that:*

$$\forall i \neq j, \quad d(R_i, R_j) \geq \frac{C_3}{\alpha}$$

   where $C_3 = \frac{1}{2}$ *ensures* $\sum_{m=k+1}^{\infty} e^{-C_3 m} \leq \epsilon$*.*

Proof. The detailed derivation of proof can be found in **Appendix** B.1. □

Table 1. Results of biomarker prediction, gene mutation prediction, and cancer subtyping tasks, with accuracy, AUC, F1 score reported. The best results are in bold, and the second-best results are underlined. Rows in gray color represent Self-Attn-based methods.

Methods	BCNB-ER ($n = 1038$)			TCGA-LUAD TP53 ($n = 469$)			UBC-OCEAN ($n = 527$)		
	ACC	AUC	F1 Score	ACC	AUC	F1 Score	ACC	AUC	F1 Score
Mean Pooling	$0.806_{\pm 0.047}$	$0.820_{\pm 0.044}$	$0.700_{\pm 0.062}$	$0.639_{\pm 0.061}$	$0.672_{\pm 0.081}$	$0.630_{\pm 0.063}$	$0.788_{\pm 0.017}$	$0.941_{\pm 0.017}$	$0.759_{\pm 0.038}$
Max Pooling	$0.810_{\pm 0.016}$	$0.819_{\pm 0.051}$	$0.723_{\pm 0.029}$	$0.659_{\pm 0.049}$	$0.660_{\pm 0.048}$	$0.650_{\pm 0.046}$	$0.800_{\pm 0.031}$	$\underline{0.946_{\pm 0.015}}$	$0.782_{\pm 0.033}$
ABMIL	$\underline{0.825_{\pm 0.038}}$	$0.825_{\pm 0.072}$	$0.726_{\pm 0.085}$	$0.639_{\pm 0.024}$	$0.688_{\pm 0.077}$	$0.630_{\pm 0.017}$	$\underline{0.823_{\pm 0.035}}$	$0.942_{\pm 0.019}$	$\underline{0.796_{\pm 0.039}}$
DS-MIL	$0.817_{\pm 0.050}$	$0.814_{\pm 0.087}$	$\underline{0.734_{\pm 0.077}}$	$0.629_{\pm 0.047}$	$0.684_{\pm 0.072}$	$0.614_{\pm 0.046}$	$0.781_{\pm 0.029}$	$0.938_{\pm 0.018}$	$0.753_{\pm 0.024}$
CLAM-SB	$0.821_{\pm 0.052}$	$0.821_{\pm 0.050}$	$0.732_{\pm 0.070}$	$0.615_{\pm 0.073}$	$0.647_{\pm 0.061}$	$0.607_{\pm 0.071}$	$0.808_{\pm 0.021}$	$0.942_{\pm 0.010}$	$\underline{0.796_{\pm 0.025}}$
DTFD	$0.786_{\pm 0.029}$	$\underline{0.835_{\pm 0.040}}$	$0.715_{\pm 0.034}$	$0.644_{\pm 0.057}$	$0.680_{\pm 0.045}$	$0.636_{\pm 0.052}$	$0.793_{\pm 0.021}$	$0.941_{\pm 0.014}$	$0.779_{\pm 0.029}$
WiKG	$0.794_{\pm 0.027}$	$0.815_{\pm 0.038}$	$0.717_{\pm 0.026}$	$0.654_{\pm 0.067}$	$0.661_{\pm 0.076}$	$0.642_{\pm 0.048}$	$0.819_{\pm 0.065}$	$0.944_{\pm 0.023}$	$0.782_{\pm 0.080}$
MambaMIL	$\underline{0.825_{\pm 0.043}}$	$0.820_{\pm 0.067}$	$0.719_{\pm 0.017}$	$0.639_{\pm 0.064}$	$0.685_{\pm 0.075}$	$0.627_{\pm 0.079}$	$0.808_{\pm 0.021}$	$0.941_{\pm 0.020}$	$0.785_{\pm 0.031}$
TransMIL	$0.802_{\pm 0.029}$	$0.780_{\pm 0.086}$	$0.664_{\pm 0.063}$	$0.615_{\pm 0.028}$	$0.658_{\pm 0.094}$	$0.596_{\pm 0.034}$	$0.723_{\pm 0.055}$	$0.928_{\pm 0.017}$	$0.698_{\pm 0.051}$
HIPT	$0.796_{\pm 0.016}$	$0.758_{\pm 0.056}$	$0.681_{\pm 0.034}$	$0.600_{\pm 0.029}$	$0.672_{\pm 0.062}$	$0.588_{\pm 0.031}$	$0.781_{\pm 0.040}$	$0.913_{\pm 0.035}$	$0.737_{\pm 0.068}$
HistGen	$0.791_{\pm 0.027}$	$0.790_{\pm 0.066}$	$0.695_{\pm 0.049}$	$\underline{0.668_{\pm 0.061}}$	$0.665_{\pm 0.056}$	$\underline{0.658_{\pm 0.063}}$	$0.784_{\pm 0.025}$	$0.934_{\pm 0.009}$	$0.764_{\pm 0.053}$
RRT-MIL	$0.812_{\pm 0.052}$	$0.814_{\pm 0.054}$	$0.688_{\pm 0.085}$	$0.600_{\pm 0.077}$	$0.657_{\pm 0.029}$	$0.581_{\pm 0.077}$	$0.804_{\pm 0.022}$	$0.939_{\pm 0.016}$	$0.783_{\pm 0.039}$
LongMIL	$0.781_{\pm 0.018}$	$0.782_{\pm 0.058}$	$0.653_{\pm 0.029}$	$0.663_{\pm 0.081}$	$\underline{0.693_{\pm 0.086}}$	$0.657_{\pm 0.079}$	$0.760_{\pm 0.022}$	$0.921_{\pm 0.016}$	$0.742_{\pm 0.033}$
Querent (Ours)	$\mathbf{0.836_{\pm 0.043}}$	$\mathbf{0.848_{\pm 0.042}}$	$\mathbf{0.739_{\pm 0.042}}$	$\mathbf{0.678_{\pm 0.068}}$	$\mathbf{0.706_{\pm 0.090}}$	$\mathbf{0.672_{\pm 0.070}}$	$\mathbf{0.835_{\pm 0.015}}$	$\mathbf{0.956_{\pm 0.019}}$	$\mathbf{0.806_{\pm 0.041}}$

3.2.4. Attentive Feature Aggregation

After obtaining contextually enhanced representations for all patches through query-aware selective attention, we employ an attentive pooling mechanism to aggregate these features for slide-level prediction. Given the refined patch features $\{x'_1, x'_2, ..., x'_N\}$, we compute attention weights through a learnable attention network:

$$w_i = \sigma(f_a(x'_i)), \quad a_i = \frac{\exp(w_i)}{\sum_{j=1}^{N} \exp(w_j)} \tag{8}$$

where f_a is a multi-layer perceptron that produces a scalar score for each patch, and σ is the sigmoid activation function. The attention weights a_i are normalized through softmax to ensure they sum to one. The final slide-level representation is computed as the weighted sum of all patch features, which is then passed through a final classification layer:

$$z = \sum_{i=1}^{N} a_i x'_i, \quad \hat{y} = f_c(z) \tag{9}$$

where f_c is a fully connected layer that outputs the predicted class probabilities. This attentive pooling mechanism allows the model to emphasize diagnostically relevant patches while suppressing the influence of irrelevant or background regions in the final prediction.

Our model is trained end-to-end using task-appropriate loss functions. For classification tasks, we employ cross-entropy (CE) loss $\mathcal{L}_{\text{ce}}$. For survival prediction tasks, we utilize the negative log-likelihood (NLL) survival loss $\mathcal{L}_{\text{surv}}$ (Cox, 1972; Katzman et al., 2018; Zadeh & Schmid, 2020). The detailed formulations of these loss functions and their optimization procedures are provided in **Appendix** C.

4. Experiments

4.1. Tasks and Datasets

We evaluate Querent across four types of tasks over 11 publicly available datasets (see **Appendix** D for details):

(1) Biomarker prediction. BCNB (Xu et al., 2021) ($n = 1038$), a dataset of early breast cancer core-needle biopsy WSI is used to predict tumor clinical characteristic estrogen receptor (ER).

(2) Gene mutation prediction. A subset of The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015) is used (TCGA-LUAD, $n = 469$) to predict TP53 gene mutation from lung adenocarcinoma WSIs.

(3) Cancer subtyping. UBC-OCEAN ($n = 527$), an ovarian cancer dataset, is used to predict 5 cancer subtypes.

(4) Survival Analysis. 8 TCGA subsets are used for this task, including TCGA-BRCA ($n = 1025$), TCGA-UCEC ($n = 497$), TCGA-STAD ($n = 365$), TCGA-LUAD ($n = 457$), TCGA-LUSC ($n = 454$), TCGA-SKCM ($n = 417$), TCGA-KIRC ($n = 500$), and TCGA-KIRP ($n = 263$).

Table 2. Results of survival prediction on 8 TCGA subsets with C-Index score reported. The best results are in bold, and the second-best results are underlined. Rows in gray color represent Self-Attn-based methods.

Methods	BRCA	UCEC	STAD	LUAD	LUSC	SKCM	KIRC	KIRP	Avg. (↑)
Mean Pooling	$0.669_{\pm 0.013}$	$0.685_{\pm 0.053}$	$0.594_{\pm 0.044}$	$0.558_{\pm 0.074}$	$0.700_{\pm 0.054}$	$0.661_{\pm 0.045}$	$0.616_{\pm 0.066}$	$0.643_{\pm 0.041}$	0.641
Max Pooling	$0.674_{\pm 0.037}$	$0.649_{\pm 0.037}$	$0.539_{\pm 0.054}$	$0.519_{\pm 0.065}$	$0.672_{\pm 0.039}$	$0.666_{\pm 0.064}$	$0.581_{\pm 0.031}$	$0.593_{\pm 0.024}$	0.612
ABMIL	$0.698_{\pm 0.020}$	$0.666_{\pm 0.039}$	$0.598_{\pm 0.075}$	$0.581_{\pm 0.063}$	$0.707_{\pm 0.042}$	$0.663_{\pm 0.084}$	$0.606_{\pm 0.065}$	$0.608_{\pm 0.071}$	0.641
DS-MIL	$0.678_{\pm 0.023}$	$0.608_{\pm 0.050}$	$0.547_{\pm 0.073}$	$0.568_{\pm 0.062}$	$0.705_{\pm 0.043}$	$0.671_{\pm 0.045}$	$0.582_{\pm 0.079}$	$0.593_{\pm 0.081}$	0.619
DTFD	$0.695_{\pm 0.022}$	$0.672_{\pm 0.032}$	$0.597_{\pm 0.090}$	$0.570_{\pm 0.054}$	$0.716_{\pm 0.046}$	$0.649_{\pm 0.048}$	$\mathbf{0.631_{\pm 0.042}}$	$0.598_{\pm 0.023}$	0.641
WiKG	$0.669_{\pm 0.019}$	$0.678_{\pm 0.049}$	$0.612_{\pm 0.066}$	$0.588_{\pm 0.056}$	$0.711_{\pm 0.030}$	$0.654_{\pm 0.083}$	$0.610_{\pm 0.031}$	$0.651_{\pm 0.036}$	0.647
MambaMIL	$0.694_{\pm 0.012}$	$0.668_{\pm 0.022}$	$0.590_{\pm 0.039}$	$\mathbf{0.609_{\pm 0.059}}$	$0.692_{\pm 0.044}$	$0.666_{\pm 0.046}$	$0.614_{\pm 0.067}$	$0.641_{\pm 0.042}$	0.647
TransMIL	$0.643_{\pm 0.038}$	$0.642_{\pm 0.049}$	$0.568_{\pm 0.082}$	$0.498_{\pm 0.067}$	$0.657_{\pm 0.056}$	$0.586_{\pm 0.042}$	$0.602_{\pm 0.043}$	$0.567_{\pm 0.092}$	0.595
HIPT	$0.717_{\pm 0.006}$	$0.647_{\pm 0.031}$	$0.588_{\pm 0.065}$	$0.536_{\pm 0.037}$	$0.603_{\pm 0.054}$	$0.616_{\pm 0.039}$	$0.578_{\pm 0.036}$	$0.559_{\pm 0.103}$	0.606
HistGen	$0.688_{\pm 0.041}$	$0.664_{\pm 0.041}$	$0.619_{\pm 0.064}$	$0.562_{\pm 0.068}$	$0.708_{\pm 0.026}$	$0.611_{\pm 0.057}$	$0.610_{\pm 0.049}$	$0.560_{\pm 0.055}$	0.628
RRT-MIL	$0.700_{\pm 0.022}$	$0.672_{\pm 0.029}$	$0.555_{\pm 0.064}$	$0.580_{\pm 0.089}$	$0.669_{\pm 0.070}$	$0.652_{\pm 0.063}$	$0.592_{\pm 0.048}$	$0.561_{\pm 0.041}$	0.623
LongMIL	$0.652_{\pm 0.024}$	$0.633_{\pm 0.062}$	$0.591_{\pm 0.094}$	$0.535_{\pm 0.038}$	$0.674_{\pm 0.074}$	$0.658_{\pm 0.026}$	$\mathbf{0.631_{\pm 0.031}}$	$0.626_{\pm 0.055}$	0.625
Querent (Ours)	$\mathbf{0.720_{\pm 0.012}}$	$\mathbf{0.691_{\pm 0.062}}$	$\mathbf{0.636_{\pm 0.028}}$	$0.608_{\pm 0.050}$	$\mathbf{0.732_{\pm 0.053}}$	$\mathbf{0.682_{\pm 0.045}}$	$0.626_{\pm 0.038}$	$\mathbf{0.667_{\pm 0.024}}$	**0.670**

These four tasks represent key challenges across the computational pathology pipeline: from molecular-level analysis (biomarker/mutation prediction) to clinical assessment (subtyping) and patient outcomes (survival).

We use 5-fold cross-validation for model training and evaluation and report the results' mean and standard deviation. For classification tasks, accuracy (ACC), area under the curve (AUC), and F1 Score are reported for comparison. For survival analysis, the concordance index (C-Index) (Harrell et al., 1982) is used for evaluation.

4.2. Baselines

We employ SOTA models including Max/Mean Pooling, ABMIL (Ilse et al., 2018), DS-MIL (Li et al., 2021a), DTFD (Zhang et al., 2022), WiKG (Li et al., 2024b), and MambaMIL (Yang et al., 2024) for comparison. For Transformer-based MIL models, we involve TransMIL (Shao et al., 2021), HIPT (Chen et al., 2022), HistGen (Guo et al., 2024), RRT-MIL (Tang et al., 2024), and LongMIL (Li et al., 2024a). For all models in comparison, we use PLIP (Huang et al., 2023) as the patch feature encoder. Further information on all baselines and implementation details are available in **Appendix** E and **Appendix** F, respectively.

5. Results

5.1. Comparison to SOTA Methods

Tab. 1 shows the evaluation results on classification tasks, including biomarker prediction (BCNB-ER), gene mutation prediction (TCGA-LUAD TP53), and cancer subtyping (UBC-OCEAN). Across all three tasks, our method consistently achieves state-of-the-art performance. For BCNB-ER, Querent achieves an accuracy of 0.836, AUC of 0.848, and F1 score of 0.739, outperforming all MIL baselines and transformer-based counterparts. On TCGA-LUAD TP53, our method demonstrates superior performance with an accuracy of 0.678 and AUC of 0.706, showing particular advantages over other transformer-based methods. For the complex task of 5-class cancer subtyping on UBC-OCEAN, Querent achieves significant improvements with an accuracy of 0.835 and AUC of 0.956, surpassing the second-best method by 1.2% and 1.0%, respectively.

Moreover, Tab. 2 demonstrates the survival prediction results across eight TCGA cancer types. Querent achieves the highest average C-Index of 0.670, representing a substantial improvement over the second-best methods (WiKG and MambaMIL, 0.647). The superior performance is consistent across different cancer types, with our method achieving the best results on BRCA (0.720), UCEC (0.691), STAD (0.636), LUSC (0.732), SKCM (0.682), and KIRP (0.667). This consistent improvement across diverse cancer types demonstrates the robustness and generalizability of our method in capturing prognostic features from WSIs.

5.2. Ablation Study

Region-level Metadata Summarization Strategy. We evaluate the effectiveness of different region-level feature summarization strategies by comparing pairwise distance rela-

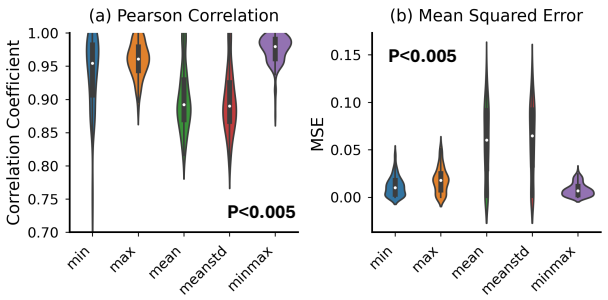

Figure 4. Ablation on **Querent** using min, max, mean, and mean ±std strategies compared to our min-max method on TCGA-LUAD TP53 gene mutation dataset (details in **Appendix** G.1).

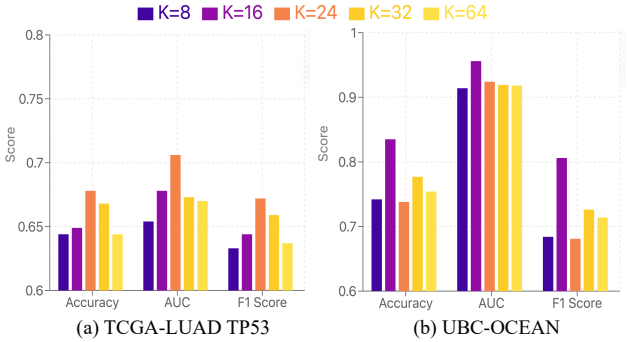

Figure 5. Ablation on **Querent** using different region size K, with reported results on TCGA-LUAD for TP53 gene mutation prediction and UBC-OCEAN for ovarian cancer subtyping.

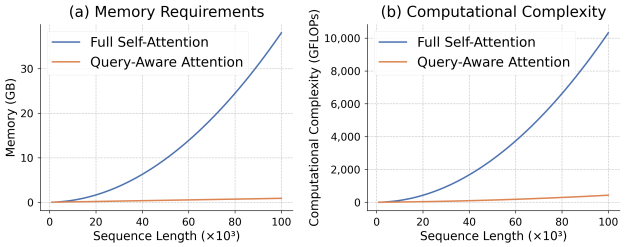

Figure 6. Computational efficiency comparison between full self-attention and our query-aware approach. (a) Memory requirements in gigabytes and (b) computational complexity in GFLOPs across different sequence lengths. See detailed analysis in **Appendix** H.

tionships between regions before and after summarization. As shown in Fig. 4, our proposed min-max summarization strategy demonstrates superior performance as measured by both the Pearson correlation coefficient and mean squared error (MSE). Specifically, our min-max approach achieves the highest correlation (0.975) and lowest MSE (0.008), significantly outperforming ($p < 0.005$) alternative methods including individual min (correlation: 0.937, MSE: 0.012) or max summarization (correlation: 0.959, MSE: 0.018), mean (correlation: 0.902, MSE: 0.058), and mean-std approaches (correlation: 0.897, MSE: 0.062), indicating that combining both minimum and maximum values effectively preserves the structural relationships between regions.

Table 3. Ablation on the importance estimation module of **Querent**, with reported results on TCGA-LUAD TP53 mutation and UBC-OCEAN ovarian cancer datasets (details in **Appendix** G.2).

Importance Estimation	Accuracy	AUC	F1 Score
TCGA-LUAD TP53 Gene Mutation Prediction			
Estimation Side Network	$0.580_{\pm 0.076}$	$0.660_{\pm 0.042}$	$0.568_{\pm 0.073}$
Random Region Selection	$0.649_{\pm 0.063}$	$0.686_{\pm 0.071}$	$0.643_{\pm 0.061}$
Querent (Ours)	$\mathbf{0.678_{\pm 0.068}}$	$\mathbf{0.706_{\pm 0.090}}$	$\mathbf{0.672_{\pm 0.070}}$
UBC-OCEAN Ovarian Cancer Subtyping			
Estimation Side Network	$0.731_{\pm 0.036}$	$0.903_{\pm 0.019}$	$0.690_{\pm 0.026}$
Random Region Selection	$0.746_{\pm 0.056}$	$0.914_{\pm 0.036}$	$0.697_{\pm 0.062}$
Querent (Ours)	$\mathbf{0.835_{\pm 0.015}}$	$\mathbf{0.956_{\pm 0.019}}$	$\mathbf{0.806_{\pm 0.041}}$

Region Importance Estimation Strategy. Meanwhile, we conduct an ablation study comparing our proposed region importance estimation module against random region selection and an estimation side network, reported in Tab. 3. Our approach consistently outperforms both baselines across TCGA-LUAD and UBC-OCEAN datasets. For TP53 mutation prediction, our method achieves 0.678 accuracy and 0.706 AUC, showing moderate improvements over the baselines. The gains are more substantial in ovarian cancer subtyping (UBC-OCEAN), where our approach reaches 0.835 accuracy and 0.956 AUC, representing absolute improvements of 8.9% in accuracy and 4.2% in AUC over random selection, demonstrating the effectiveness of our region importance estimation strategy.

Region size in Querent. Fig. 5 shows the impact of applying different region size for our Querent model during the region metadata summarization process. For TP53 mutation prediction in TCGA-LUAD, moderate region sizes (K=24) yielded optimal results with an AUC of 0.706, while both smaller (K=8) and larger (K=64) regions showed decreased performance. In UBC-OCEAN ovarian cancer subtyping, K=16 emerged as the clear optimal choice, achieving the highest accuracy (0.835) and AUC (0.956), with performance gradually declining as region size increased. These results demonstrate that moderate-sized regions are most effective for both tasks. This aligns with pathological intuition, as larger regions may dilute the discriminative local tissue patterns by aggregating potentially heterogeneous areas, while smaller regions might miss important contextual information.

5.3. Computational Efficiency

We analyze the computational efficiency of our query-aware attention mechanism compared to full self-attention (illus-

trated in Fig. 6). While full self-attention reaches 37GB memory and 10,000 GFLOPs at 100k sequence length with quadratic scaling, our approach achieves near-linear scaling, requiring only 1GB memory and 500 GFLOPs for a 100k patch sequence. This significant reduction in computational overhead enables our model to efficiently process longer sequences, which is crucial for gigapixel whole-slide image analysis.

6. Conclusion

In this study, we present **Querent**, a query-aware dynamic modeling framework for understanding long-range contextual correlations in gigapixel WSIs. Inspired by pathological observations, our method aims to dynamically adapt to the unique context of each patch while remaining computationally efficient through region-level metadata summarization and importance-based attention. Extensive experiments across multiple CPath tasks demonstrate the superior performance of our model compared to state-of-the-art approaches, establishing its effectiveness in whole slide image analysis.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62202403), Hong Kong Innovation and Technology Commission (Project No. MHP/002/22 and ITCPD/17-9), and Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. R6003-22 and C4024-22GF).

Impact Statement

This work advances computational pathology through a novel query-aware dynamic long contextual modeling framework for analyzing gigapixel whole slide images. Our research utilizes publicly available pathology datasets with appropriate institutional approvals. While our work demonstrates potential for improving the efficacy and efficiency of histopathological analysis, it is currently intended for research purposes only. Clinical implementation would require extensive external validation and regulatory approval. Our methodology and findings are shared to advance scientific understanding in computational pathology rather than for immediate clinical application.

References

Araujo, A., Norris, W., and Sim, J. Computing receptive fields of convolutional neural networks. *Distill*, 4(11): e21, 2019.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8): 1301–1309, 2019.

Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., and Mahmood, F. 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, 2022.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3): 850–862, 2024.

Chikontwe, P., Kim, M., Nam, S. J., Go, H., and Park, S. H. Multiple instance learning with center embeddings for histopathology classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pp. 519–528. Springer, 2020.

Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Cui, M. and Zhang, D. Y. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4): 412–422, 2021.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Du, Y., Zhuang, J., Zheng, X., Cong, J., Guo, L., He, C., Luo, L., and Li, X. Beyond h&e: Unlocking pathological insights with polarization via self-supervised learning. *arXiv preprint arXiv:2503.05933*, 2025.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Guo, Z., Ma, J., Xu, Y., Wang, Y., Wang, L., and Chen, H. 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*, pp. 189–199. Springer, 2024.

Guo, Z., Xiong, C., Ma, J., Sun, Q., Feng, L., Wang, J., and Chen, H. Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Han, D., Pan, X., Han, Y., Song, S., and Huang, G. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5961–5971, 2023.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

Heindl, A., Nawaz, S., and Yuan, Y. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. *Laboratory investigation*, 95(4): 377–384, 2015.

Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., and Wang, L. Hˆ 2-mil: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 933–941, 2022.

Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J., and Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

Ilse, M., Tomczak, J., and Welling, M. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Jin, C., Guo, Z., Lin, Y., Luo, L., and Chen, H. Label-efficient deep learning in medical image analysis: Challenges and future directions. *arXiv preprint arXiv:2303.12484*, 2023.

Kaban, A. Improved bounds on the dot product under random projection and random sign projection. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 487–496, 2015.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18: 1–12, 2018.

Li, B., Li, Y., and Eliceiri, K. W. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14318–14328, 2021a.

Li, H., Yang, F., Zhao, Y., Xing, X., Zhang, J., Gao, M., Huang, J., Wang, L., and Yao, J. Dt-mil: deformable transformer for multi-instance learning on histopathological image. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pp. 206–216. Springer, 2021b.

Li, H., Yang, L., Wang, Y., Wang, L., Chen, G., Zhang, L., and Wang, D. Integrative analysis of tp53 mutations in lung adenocarcinoma for immunotherapies and prognosis. *BMC bioinformatics*, 24(1):155, 2023.

Li, H., Zhang, Y., Chen, P., Shui, Z., Zhu, C., and Yang, L. Rethinking transformer for long contextual histopathology whole slide image analysis. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 101498–101528. Curran Associates, Inc., 2024a.

Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., and He, Y. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11323–11332, 2024b.

Lin, T., Yu, Z., Hu, H., Xu, Y., and Chen, C.-W. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19830–19839, 2023.

Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Lu, M. Y., Chen, B., Zhang, A., Williamson, D. F., Chen, R. J., Ding, T., Le, L. P., Chuang, Y.-S., and Mahmood, F. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19764–19775, 2023.

Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L. P., Gerber, G., et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

Ma, J., Guo, Z., Zhou, F., Wang, Y., Xu, Y., Cai, Y., Zhu, Z., Jin, C., Jiang, Y. L. X., Han, A., et al. Towards a generalizable pathology foundation model via unified knowledge distillation. *arXiv preprint arXiv:2407.18449*, 2024.

Niazi, M. K. K., Parwani, A. V., and Gurcan, M. N. Digital pathology and artificial intelligence. *The lancet oncology*, 20(5):e253–e261, 2019.

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. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Song, A. H., Jaume, G., Williamson, D. F., Lu, M. Y., Vaidya, A., Miller, T. R., and Mahmood, F. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, 2023.

Song, A. H., Chen, R. J., Ding, T., Williamson, D. F., Jaume, G., and Mahmood, F. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11566–11578, 2024a.

Song, A. H., Chen, R. J., Jaume, G., Vaidya, A. J., Baras, A., and Mahmood, F. Multimodal prototyping for cancer survival prediction. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46050–46073. PMLR, 21–27 Jul 2024b.

Sun, Q., Guo, Z., Peng, R., Chen, H., and Wang, J. Any-to-any learning in computational pathology via triplet multimodal pretraining. *arXiv preprint arXiv:2505.12711*, 2025.

Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., and Liu, B. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11343–11352, 2024.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

Van der Laak, J., Litjens, G., and Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.

Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Wang, W., Ma, S., Xu, H., Usuyama, N., Ding, J., Poon, H., and Wei, F. When an image is worth 1,024 x 1,024 words: A case study in computational pathology. *arXiv preprint arXiv:2312.03558*, 2023.

Wang, X., Xiang, J., Zhang, J., Yang, S., Yang, Z., Wang, M.-H., Zhang, J., Yang, W., Huang, J., and Han, X. Scl-wc: Cross-slide contrastive learning for weakly-supervised whole-slide image classification. *Advances in neural information processing systems*, 35:18009–18021, 2022.

Xiang, T., Song, Y., Zhang, C., Liu, D., Chen, M., Zhang, F., Huang, H., O'Donnell, L., and Cai, W. Dsnet: A dual-stream framework for weakly-supervised gigapixel pathology image analysis. *IEEE Transactions on Medical Imaging*, 41(8):2180–2190, 2022.

Xiong, C., Chen, H., Sung, J. J., and King, I. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. *arXiv preprint arXiv:2301.08125*, 2023.

Xiong, C., Chen, H., Zheng, H., Wei, D., Zheng, Y., Sung, J. J., and King, I. Mome: Mixture of multimodal experts for cancer survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 318–328. Springer, 2024a.

Xiong, C., Lin, Y., Chen, H., Zheng, H., Wei, D., Zheng, Y., Sung, J. J., and King, I. Takt: Target-aware knowledge transfer for whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 503–513. Springer, 2024b.

Xiong, C., Chen, H., and Sung, J. J. A survey of pathology foundation model: Progress and future directions. *arXiv preprint arXiv:2504.04045*, 2025a.

Xiong, C., Guo, Z., Xu, Z., Zhang, Y., Tong, R. K.-Y., Yeo, S. Y., Chen, H., Sung, J. J., and King, I. Beyond linearity: Squeeze-and-recalibrate blocks for few-shot whole slide image classification. *arXiv preprint arXiv:2505.15504*, 2025b.

Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., and Singh, V. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14138–14148, 2021.

Xu, F., Zhu, C., Tang, W., Wang, Y., Zhang, Y., Li, J., Jiang, H., Shi, Z., Liu, J., and Jin, M. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in Oncology*, pp. 4133, 2021.

Xu, Y. and Chen, H. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21241–21251, 2023.

Xu, Y., Wang, Y., Zhou, F., Ma, J., Yang, S., Lin, H., Wang, X., Wang, J., Liang, L., Han, A., et al. A multimodal knowledge-enhanced whole-slide pathology foundation model. *arXiv preprint arXiv:2407.15362*, 2024.

Yang, S., Wang, Y., and Chen, H. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 296–306. Springer, 2024.

Yu, J.-G., Wu, Z., Ming, Y., Deng, S., Li, Y., Ou, C., He, C., Wang, B., Zhang, P., and Wang, Y. Prototypical multiple instance learning for predicting lymph node metastasis of breast cancer from whole-slide pathological images. *Medical Image Analysis*, 85:102748, 2023.

Yuan, Y. Spatial heterogeneity in the tumor microenvironment. *Cold Spring Harbor perspectives in medicine*, 6 (8):a026583, 2016.

Zadeh, S. G. and Schmid, M. Bias in cross-entropy-based training of deep survival networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3126–3137, 2020.

Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S. E., and Zheng, Y. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18802–18812, 2022.

Zheng, X., Wen, J., Zhuang, J., Du, Y., Cong, J., Guo, L., He, C., Luo, L., and Chen, H. Diffusion-based virtual staining from polarimetric mueller matrix imaging. *arXiv preprint arXiv:2503.01352*, 2025.

Zheng, Y., Gindra, R. H., Green, E. J., Burks, E. J., Betke, M., Beane, J. E., and Kolachalama, V. B. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022.

Zhou, F. and Chen, H. Cross-modal translation and alignment for survival analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21485–21494, 2023.

A. Algorithm Pseudo Code of Querent

Algorithm 1 Querent: Query-Aware Dynamic Long Sequence Modeling for Gigapixel WSI Analysis

Input: Patch features $X = \{x_1, x_2, ..., x_N\}$, region size K, number of regions M
Output: Slide-level prediction $\hat{y}$ and attention weights α
Parameter: Query projection f_q, metadata projections $f_{\text{min}}, f_{\text{max}}$, attention $\mathbf{W}_{\text{qkv}}, \mathbf{W}_O$, classifier f_c
// Phase 1: Region-level Metadata Summarization
for $i = 1$ to M **do**
  Extract region patches: $R_i = \{x_{i1}, x_{i2}, ..., x_{iK}\}$
  Compute min/max metadata: $m_i^{\text{min}} = \min_{j\in\{1...K\}} x_{ij}$, $m_i^{\text{max}} = \max_{j\in\{1...K\}} x_{ij}$
  Project metadata: $\hat{m}_i^{\text{min}} = f_{\text{min}}(m_i^{\text{min}})$, $\hat{m}_i^{\text{max}} = f_{\text{max}}(m_i^{\text{max}})$
end for
// Phase 2: Region Importance Estimation
for each query patch q **do**
  Project query: $\hat{q} = f_q(q)$
  for $i = 1$ to M **do**
    Compute importance: $s_i = \max(|\langle \hat{q}, \hat{m}_i^{\text{min}} \rangle|, |\langle \hat{q}, \hat{m}_i^{\text{max}} \rangle|)$
  end for
  Select top regions: $\mathcal{R}_q = \text{TopK}(\{(R_i, s_i)\}_{i=1}^{M})$
end for
// Phase 3: Query-Aware Selective Attention
for each selected region $R_i \in \mathcal{R}_q$ **do**
  Compute QKV: $[\mathbf{Q}, \mathbf{K}, \mathbf{V}] = \mathbf{W}_{\text{qkv}}(x)$
  Compute attention scores: $\mathbf{A} = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_h}})$
  Compute attention output: $\mathbf{O} = \mathbf{A}\mathbf{V}$
  Multi-head attention: $\mathbf{O}_h = \text{Concat}(\mathbf{O}_1, \ldots, \mathbf{O}_H)\mathbf{W}_O$
end for
// Phase 4: Attentive Feature Aggregation
Compute attention weights: $w_i = \sigma(f_a(x_i'))$
Normalize weights: $a_i = \frac{\exp(w_i)}{\sum_{j=1}^{N} \exp(w_j)}$
Aggregate features: $z = \sum_{i=1}^{N} a_i x_i'$
Compute final prediction: $\hat{y} = f_c(z)$
Return $\hat{y}$, α

B. Theoretical Analysis

B.1. Query-Aware Sparse Attention: A Theoretically Bounded Approximation of Full Self-Attention

B.1.1. Definitions and Preliminaries

Definition B.1 (Full Self-Attention Matrix)**.** Let $\mathbf{X} \in \mathbb{R}^{L\times d}$ be the feature matrix of a sequence of patches, where L is the sequence length and d is the feature dimension. The full self-attention matrix $\mathbf{B}$ is defined as:

$$\mathbf{B}_{i,j} = \exp\left(\frac{\mathbf{q}_i^\top \mathbf{K}_j}{\sqrt{d}}\right)$$

where $\mathbf{q}_i$ is the query vector for patch i and $\mathbf{K}_j$ is the key vector for patch j.

Definition B.2 (Query-Aware Attention Matrix)**.** Let $\mathbf{X} \in \mathbb{R}^{L\times d}$ be the feature matrix of a sequence of patches, where L is the sequence length and d is the feature dimension. The query-aware attention matrix $\mathbf{A}$ is defined as:

$$\mathbf{A}_{i,j} = \begin{cases} \exp\left(\frac{\mathbf{q}_i^\top \mathbf{K}_j}{\sqrt{d}}\right) & \text{if } j \in \mathcal{R}_i \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{R}_i$ is the set of top-K relevant regions for query patch $\mathbf{q}_i$, and $\mathbf{K}_j$ is the key vector for patch j.

Definition B.3 (Region-Level Metadata)**.** For each region R_i, the metadata consists of summary statistics computed from the patches within the region. The primary statistics are the minimum and maximum feature vectors:

$$\mathbf{m}^i_{\min} = \min_{j \in R_i} \mathbf{x}_j, \quad \mathbf{m}^i_{\max} = \max_{j \in R_i} \mathbf{x}_j$$

These vectors are projected via Lipschitz-continuous neural networks $f_{\min}$ and $f_{\max}$ into a shared embedding space:

$$\hat{\mathbf{m}}^i_{\min} = f_{\min}(\mathbf{m}^i_{\min}), \quad \hat{\mathbf{m}}^i_{\max} = f_{\max}(\mathbf{m}^i_{\max})$$

where $f_{\min}$ and $f_{\max}$ are L-Lipschitz continuous with constant L.

B.1.2. TECHNICAL LEMMAS

Lemma B.4 (Region Metadata Approximation)**.** *Let B be a bound on the input norms $\|\mathbf{q}_i\|, \|\mathbf{K}_j\| \leq B$. For any query patch $\mathbf{q}$ and region R_i, assuming L-Lipschitz continuous projection functions, the interaction scores satisfy:*

$$\max_{j \in R_i} |\langle \mathbf{q}, \mathbf{x}_j \rangle - \langle \hat{\mathbf{q}}, \hat{\mathbf{m}}^i_{\min} \rangle| \leq B \cdot L \cdot \operatorname{diam}(R_i) = \epsilon_1$$

and similarly for $\hat{\mathbf{m}}^i_{\max}$, where $\operatorname{diam}(R_i)$ is the diameter of region R_i in feature space.

Proof. Fix any patch $\mathbf{x}_j \in R_i$. By construction:

$$\mathbf{m}^i_{\min} \preceq \mathbf{x}_j \preceq \mathbf{m}^i_{\max}$$

where $\preceq$ denotes element-wise comparison. By Lipschitz continuity:

$$\|\hat{\mathbf{m}}^i_{\min} - f_{\min}(\mathbf{x}_j)\| \leq L \|\mathbf{m}^i_{\min} - \mathbf{x}_j\| \leq L \cdot \operatorname{diam}(R_i)$$

Using Cauchy-Schwarz and the norm bound B:

$$|\langle \mathbf{q}, \mathbf{x}_j \rangle - \langle \hat{\mathbf{q}}, \hat{\mathbf{m}}^i_{\min} \rangle| \leq \|\mathbf{q}\| \cdot L \cdot \operatorname{diam}(R_i) \leq B \cdot L \cdot \operatorname{diam}(R_i) = \epsilon_1.$$

□

Lemma B.5 (Ranking Stability)**.** *Let $s_i = \max(|\langle \hat{\mathbf{q}}, \hat{\mathbf{m}}^i_{\min} \rangle|, |\langle \hat{\mathbf{q}}, \hat{\mathbf{m}}^i_{\max} \rangle|)$. If $\max_{j \in R_i} |\langle \mathbf{q}, \mathbf{x}_j \rangle - \langle \hat{\mathbf{q}}, \hat{\mathbf{m}}^i_{\min} \rangle| \leq \epsilon_1$, then the top-$K$ regions selected by s_i and true interactions differ by at most $2\epsilon_1$-suboptimal regions.*

Proof of Lemma B.5. Let S^K_{true} denote the true top-K regions ranked by $\langle \mathbf{q}, \mathbf{x}_j \rangle$, and S^K_{approx} denote those selected by s_i. For any region $R_i \in S^K_{\text{approx}} \setminus S^K_{\text{true}}$, its approximate score satisfies:

$$s_i \geq \min_{R \in S^K_{\text{approx}}} s_R \geq \max_{R \notin S^K_{\text{approx}}} s_R.$$

By the approximation error bound ϵ_1, we have:

$$\langle \mathbf{q}, \mathbf{x}_j \rangle \geq s_i - \epsilon_1 \quad \text{and} \quad \langle \mathbf{q}, \mathbf{x}_j \rangle \leq s_i + \epsilon_1 \quad \forall j \in R_i.$$

Thus, any region $R_i \in S^K_{\text{approx}}$ must satisfy:

$$\langle \mathbf{q}, \mathbf{x}_j \rangle \geq \left(\max_{R \notin S^K_{\text{approx}}} s_R \right) - \epsilon_1.$$

This implies R_i is at most $2\epsilon_1$-suboptimal compared to the true top-K regions. □

B.1.3. MAIN THEOREM

Theorem B.6 (Query-Aware Attention Approximation)**.** *Let* $\mathbf{A}$ *be the query-aware attention matrix (Def. B.2), and* $\mathbf{B}$ *be the full self-attention matrix (Def. B.1). Assume attention scores decay exponentially with spatial distance:* $\exp(-\alpha d(i,j))$ *bounds the attention score decay for distance* $d(i,j)$*. For any input sequence* $\mathbf{X}$ *with* L *patches and* d *dimensions, there exist* ***random projection matrices*** $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d\times d}$ *such that:*

$$\|\mathbf{A} - \mathbf{B}\|_F \leq \left(2 + \frac{B}{\sqrt{d}}\right)\epsilon$$

with probability at least $1-\delta$*, provided:*

1. *The hidden dimension satisfies:*
$$d \geq C_1 \cdot \frac{\log(L/\delta)}{\epsilon^2}$$
where $C_1 = 8B^4$ *(from JL inner-product preservation (Kaban, 2015)).*

2. *The number of selected regions per query satisfies:*
$$k \geq \frac{C_2}{\alpha} \cdot \log\left(\frac{1}{\epsilon}\right)$$
where $C_2 = 2$ *(derived from Step 3).*

3. *For each region* R_i*, the diameter satisfies:*
$$\operatorname{diam}(R_i) \leq \min\left(\frac{\epsilon}{L\cdot\sqrt{d}}, \frac{1}{\alpha}\right)$$

4. *Regions are spatially separated such that:*
$$\forall i \neq j, \quad d(R_i, R_j) \geq \frac{C_3}{\alpha}$$
where $C_3 = \frac{1}{2}$ *ensures* $\sum_{m=k+1}^{\infty} e^{-C_3 m} \leq \epsilon$*.*

Proof of Theorem B.6.

Step 1: Region Metadata Summarization.
By Lemma B.4, projections $\hat{\mathbf{m}}^i_{\min}, \hat{\mathbf{m}}^i_{\max}$ approximate interactions within R_i with error $\epsilon_1 = B \cdot L \cdot \operatorname{diam}(R_i) \leq \frac{B\epsilon}{\sqrt{d}}$. The condition $\operatorname{diam}(R_i) \leq \frac{1}{\alpha}$ ensures region-level interactions respect the exponential decay rate α.

Step 2: Region Importance Estimation.
Using Lemma B.5, approximate scores s_i preserve the true top-K regions up to error ϵ_1. Thus, $\mathcal{R}_i$ includes all regions with $s_i \geq \max_j s_j - 2\epsilon_1$.

Step 3: Bounding Non-Selected Regions.
Under spatial separation $d(R_i, R_j) \geq \frac{C_3}{\alpha}$, the tail sum becomes:

$$\sum_{j\notin\mathcal{R}_i} \exp(-\alpha d(i,j)) \leq \sum_{m=k+1}^{\infty} e^{-C_3 m} = \frac{e^{-C_3(k+1)}}{1-e^{-C_3}} \leq \epsilon$$

Solving for k gives $k \geq \frac{1}{C_3}\log\left(\frac{1}{\epsilon(1-e^{-C_3})}\right)$. Setting $C_3 = \frac{1}{2}$ simplifies this to $k \geq \frac{2}{\alpha}\log\left(\frac{1}{\epsilon}\right)$ (i.e., $C_2 = 2$).

Step 4: Johnson-Lindenstrauss Guarantee.
Using random matrices $\mathbf{W}_Q, \mathbf{W}_K$ (Kaban, 2015), for $d \geq 8B^4\log(L/\delta)/\epsilon^2$, we have:

$$\Pr\left[\left|\langle\hat{\mathbf{q}}, \hat{\mathbf{K}}_j\rangle - \langle\mathbf{q}, \mathbf{K}_j\rangle\right| \leq \epsilon\right] \geq 1-\delta$$

Step 5: Frobenius Norm Aggregation.
Normalize $\epsilon \leftarrow \epsilon/L$ to absorb the L^2 scaling. The entrywise error becomes:

$$\|\mathbf{A}-\mathbf{B}\|_F \leq \sqrt{L^2\left(2\epsilon+\frac{B\epsilon}{\sqrt{d}}\right)^2} = L\left(2\epsilon+\frac{B\epsilon}{\sqrt{d}}\right) \xrightarrow{\epsilon\leftarrow\epsilon/L} \left(2+\frac{B}{\sqrt{d}}\right)\epsilon$$

□

B.1.4. DISCUSSION

The theorem shows query-aware attention effectively approximates full attention via region metadata, sufficient embedding dimension, and region selection. Error bounds depend on L, d, B, k, α, and region diameters. The spatial separation condition ensures the exponential decay property is meaningful.

C. Task-specific Losses

C.1. Cross-entropy Loss for Classification Tasks

For classification tasks (biomarker prediction, gene mutation prediction, and cancer subtyping), we employ the standard cross-entropy (CE) loss. Given a WSI X with slide-level label y, and model prediction $\hat{y} = f(X;\theta)$ where $\hat{y} \in \mathbb{R}^C$ represents the predicted probabilities over C classes, the CE loss is defined as:

$$\mathcal{L}_{ce} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{10}$$

where y_c is the one-hot encoded ground truth label for class c. The final classification loss is averaged over all WSIs in the training set:

$$\mathcal{L}_{cls} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{ce}^{(i)} \tag{11}$$

where N is the number of WSIs in the training set and $\mathcal{L}_{ce}^{(i)}$ is the CE loss for the i-th WSI.

This loss function encourages the model to output probability distributions that assign high probabilities to the correct classes while minimizing probabilities for incorrect classes, effectively training the query-aware attention mechanism to focus on diagnostically relevant regions within each WSI.

C.2. Negative Log-likelihood (NLL) Loss for Survival Analysis

The NLL survival loss (Zadeh & Schmid, 2020; Song et al., 2024b) generalizes the standard negative log-likelihood loss to accommodate censored data. The goal is to predict patient survival based on their patient-level embedding, $\bar{x}_{\text{patient}} \in \mathbb{R}^{2d}$. For each patient, we consider two key pieces of information: (1) a censorship status c, where $c = 0$ indicates observed death and $c = 1$ indicates the last known follow-up, and (2) a time-to-event t_i, representing either time until death ($c = 0$) or time until last follow-up ($c = 1$).

Rather than directly predicting t_i, we discretize time into n non-overlapping intervals (t_{j-1}, t_j), where $j \in [1, ..., n]$, based on the quartiles of observed survival times, and denote each interval as y_j. This transforms survival prediction into a classification problem, where each patient's outcome is defined by $(\bar{x}_{\text{patient}}, y_j, c)$.

For each interval, we compute a hazard function $f_{\text{hazard}}(y_j|\bar{x}_{\text{patient}}) = S(\hat{y}_j)$, where S is the sigmoid activation. Intuitively, $f_{\text{hazard}}(y_j|\bar{x}_{\text{patient}})$ represents the probability of death occurring within interval (t_{j-1}, t_j). We also define a survival function $f_{\text{surv}}(y_j|\bar{x}_{\text{patient}}) = \prod_{k=1}^{j}(1 - f_{\text{hazard}}(y_k|\bar{x}_{\text{patient}}))$ that represents the probability of survival up to interval (t_{j-1}, t_j).

The NLL survival loss for a dataset of N_D patients can then be formalized as:

$$\mathcal{L}\left(\{\bar{x}^{(i)}_{\text{patient}}, y^{(i)}_j, c^{(i)}\}_{i=1}^{N_D}\right) = \sum_{i=1}^{N_D} -c^{(i)} \log(f_{\text{surv}}(y^{(i)}_j|\bar{x}^{(i)}_{\text{patient}})) \\ - (1-c^{(i)}) \log(f_{\text{surv}}(y^{(i)}_{j-1}|\bar{x}^{(i)}_{\text{patient}})) - (1-c^{(i)}) \log(f_{\text{hazard}}(y^{(i)}_j|\bar{x}^{(i)}_{\text{patient}})) \tag{12}$$

The first term enforces a high survival probability for patients still alive at their last follow-up. The second term ensures high survival probability up to the time interval before death for uncensored patients. The third term promotes accurate prediction of the death interval for uncensored patients.

D. Datasets

BCNB[1] (Xu et al., 2021) for biomarker prediction. Short for Early Breast Cancer Core-Needle Biopsy WSI, the BCNB dataset includes core-needle biopsy WSIs of early breast cancer patients and the corresponding clinical data. 1038 WSIs are annotated with their corresponding estrogen receptor (ER) expression, which is a favorable prognostic parameter in breast cancer, and a predictor for response to endocrine therapy.

TCGA-LUAD[2] (Tomczak et al., 2015) for TP53 gene mutation prediction. TCGA-LUAD TP53 dataset includes 469 lung adenocarcinoma WSIs with their corresponding annotation of mutation of the TP53 tumor suppressor gene, which is one of the most mutated genes in lung adenocarcinoma and represents a vital role in regulating the occurrence and progression of cancer (Li et al., 2023).

UBC-OCEAN[3] for ovarian cancer subtyping. UBC-OCEAN is a large-scale ovarian cancer dataset collected from over 20 medical centers across four continents. The dataset collects 527 ovarian cancer WSIs, representing five major ovarian carcinoma subtypes: high-grade serous carcinoma, clear-cell carcinoma, endometrioid, low-grade serous, and mucinous carcinoma, along with several rare subtypes. Each subtype exhibits distinct morphological patterns, molecular profiles, and clinical characteristics, making accurate subtype classification crucial for treatment planning. The challenge of accurate subtype identification is particularly relevant given the growing emphasis on subtype-specific treatment approaches and the limited availability of specialist gynecologic pathologists in many regions.

TCGA Subsets[4] (Tomczak et al., 2015) for survival analysis. In this work, we include 8 TCGA sub-datasets for unimodal survival prediction:

1. TCGA-BRCA: Breast Invasive Carcinoma dataset from TCGA, containing 1025 whole slide images with associated survival outcome data for breast cancer patients. This comprehensive collection represents one of the largest breast cancer datasets, featuring diverse histological patterns and molecular subtypes including ductal and lobular carcinomas. The survival data encompasses overall survival time and vital status, enabling robust prognostic model development.
2. TCGA-UCEC: Uterine Corpus Endometrial Carcinoma dataset from TCGA, encompassing 497 WSIs and survival data from endometrial cancer cases. This dataset includes various histological grades and stages of endometrial carcinoma, providing valuable insights into the progression and prognosis of gynecologic cancers. The images showcase diverse morphological patterns characteristic of endometrial malignancies.
3. TCGA-STAD: Stomach Adenocarcinoma dataset from TCGA, which includes 365 WSIs and patient survival information for gastric cancer cases. The dataset represents different anatomical locations within the stomach and various histological subtypes of gastric adenocarcinoma. The survival data is particularly valuable for understanding the relationship between morphological features and patient outcomes in gastric cancer.
4. TCGA-LUAD: Lung Adenocarcinoma dataset from TCGA, containing 457 WSIs and survival outcome data from lung adenocarcinoma patients. This collection captures the heterogeneous nature of lung adenocarcinomas, including various growth patterns and degrees of differentiation. The survival information is crucial for developing predictive models for one of the most common types of lung cancer.

[1] https://bcnb.grand-challenge.org/
[2] https://portal.gdc.cancer.gov/
[3] https://www.kaggle.com/competitions/UBC-OCEAN
[4] https://portal.gdc.cancer.gov/

5. TCGA-LUSC: Lung Squamous Cell Carcinoma dataset from TCGA, providing 454 WSIs and survival data for lung squamous cell carcinoma cases. These images showcase the distinct morphological features of squamous cell carcinomas, including keratinization and intercellular bridges. The dataset enables comparative studies between different types of lung cancers and their prognostic factors.
6. TCGA-SKCM: Skin Cutaneous Melanoma dataset from TCGA, consisting of 417 WSIs and survival information from melanoma patients. This collection includes primary and metastatic melanoma cases, capturing the diverse histological patterns and progression stages of this aggressive skin cancer. The survival data is particularly relevant for understanding metastatic potential and treatment response.
7. TCGA-KIRC: Kidney Renal Clear Cell Carcinoma dataset from TCGA, containing 500 WSIs with survival outcome data for kidney cancer patients. The dataset showcases the characteristic clear cell morphology and various grades of renal cell carcinoma. The survival information helps in understanding the prognostic implications of morphological variations in kidney cancer.
8. TCGA-KIRP: Kidney Renal Papillary Cell Carcinoma dataset from TCGA, providing 263 WSIs and survival data for papillary renal cell carcinoma cases. This collection represents a distinct histological subtype of kidney cancer, featuring papillary architecture and different cellular patterns. The survival data enables comparative analysis with other renal cancer subtypes.

All these datasets are part of The Cancer Genome Atlas (TCGA) program and contain high-resolution histopathology whole slide images along with corresponding patient survival data. Each dataset has been digitized following standardized protocols and includes detailed clinical annotations. The images are typically scanned at 40x magnification, providing high-detail visualization of cellular and architectural features. These comprehensive resources enable researchers to develop and validate computational methods for survival prediction based on histopathological features, facilitating advances in precision oncology and personalized medicine approaches.

E. Baselines

Max/Mean Pooling. Mean Pooling represents one of the most straightforward aggregation strategies in multi-instance learning. This approach processes all instances within a bag equally, combining their features through average pooling to create a single, unified representation. While simple in nature, this method effectively captures the collective characteristics of all instances, making it particularly suitable for cases where every instance contributes meaningful information to the overall bag classification. Max Pooling, on the other hand, adopts a more selective approach by identifying and utilizing the most prominent features across all instances. This strategy operates under the assumption that the most distinctive or highest-activated features are the most relevant for classification. By focusing on these peak features, max pooling can effectively highlight the most discriminative patterns within the bag, though it may potentially overlook more subtle, collective patterns that could be captured by other methods.

ABMIL (Ilse et al., 2018). Attention-Based Multiple Instance Learning (ABMIL) enhances bag-level feature aggregation by incorporating a learnable attention mechanism that adaptively weights the importance of different instances within a bag. Unlike fixed pooling strategies, ABMIL computes attention scores for each instance based on their learned representations, enabling the model to emphasize the most informative elements while downweighting less relevant ones. This adaptive weighting scheme, combined with its inherent interpretability through attention weights, makes ABMIL particularly effective for scenarios where instances have varying levels of relevance to the classification task.

DS-MIL (Li et al., 2021a). Dual-Stream Multiple Instance Learning (DS-MIL) advances bag-level feature aggregation through a novel two-stream architecture that combines the benefits of instance-level and bag-level learning. The model's first stream employs max pooling to identify the most discriminative instance (critical instance), while the second stream computes attention weights for all instances based on their learned distance to the critical instance. By fusing these complementary streams and incorporating trainable distance measurements, DS-MIL creates a more nuanced decision boundary that better captures the relationships between instances, making it particularly effective for scenarios with complex instance distributions within positive bags.

DTFD (Zhang et al., 2022). DTFD introduces an innovative dual-tier framework for whole slide image classification that addresses the inherent challenges of limited training samples with high instance counts. The model's key innovation lies in its pseudo-bag strategy, which virtually increases the number of training bags by randomly splitting instances from each

original slide into smaller subsets. A two-tier architecture processes these pseudo-bags: the first tier applies attention-based MIL to the pseudo-bags, while the second tier distills features from the first tier's outputs to generate final slide-level predictions. This hierarchical approach, combined with multiple feature distillation strategies, enables more effective learning from limited training data while maintaining robustness against noise in positive instance distributions.

WiKG (Li et al., 2024b). WiKG introduces a dynamic graph representation framework for whole slide image analysis that conceptualizes the relationships between image patches through a knowledge graph structure. The model employs a dual-stream approach: head embeddings that explore correlations between patches and tail embeddings that capture each patch's contribution to others. These embeddings are then used to dynamically construct directed edges between patches based on their learned relationships. A knowledge-aware attention mechanism aggregates information across patches by computing attention scores through the joint modeling of head, tail, and edge embeddings. This architecture allows WiKG to both capture long-range dependencies between distant patches and model directional information flow, overcoming key limitations of conventional graph-based and instance-based approaches for histopathology image analysis.

MambaMIL (Yang et al., 2024). MambaMIL introduces a novel approach to whole slide image analysis by incorporating Selective Scan Space State Sequential Model, *i.e.*, Mamba (Gu & Dao, 2023), into multiple instance learning with linear complexity. The model's core innovation lies in its Sequence Reordering Mamba (SR-Mamba) architecture, which processes instance sequences in dual streams - one preserving the original sequence order and another utilizing reordered sequences. This dual-stream approach enables more comprehensive modeling of relationships between instances while maintaining computational efficiency. The SR-Mamba module dynamically reorders instance sequences within non-overlapping segments, allowing the model to capture both local and global dependencies between patches. Through this sequence reordering mechanism and linear-time sequence modeling inherited from Mamba, the model effectively addresses common challenges in whole slide image analysis such as overfitting and high computational overhead.

TransMIL (Shao et al., 2021). TransMIL introduces a novel correlated multiple instance learning framework that fundamentally reimagines how relationships between instances are modeled in whole slide image analysis. Unlike traditional MIL approaches that assume instances are independent and identically distributed, TransMIL leverages Transformer architecture to capture both morphological and spatial correlations between patches. The model employs a TPT (Transformer Pyramid Translation) module that combines multi-head self-attention for modeling instance relationships with position-aware encoding to preserve spatial context.

HIPT (Chen et al., 2022). HIPT (Hierarchical Image Pyramid Transformer) is a novel architecture designed specifically for analyzing gigapixel whole-slide images in computational pathology. It processes images in a hierarchical manner across multiple scales - from cell-level features (16×16 pixels) to larger tissue regions (4096×4096 pixels) - mirroring how pathologists examine slides by zooming in and out. This hierarchical approach allows HIPT to capture both fine-grained cellular details and broader tissue patterns simultaneously. The model employs a series of Vision Transformers arranged in a pyramid structure to aggregate information across these different scales, enabling it to learn representations that incorporate both local and global tissue contexts.

HistGen (Guo et al., 2024). HistGen is a hierarchical architecture designed for efficient processing and representation learning of gigapixel whole slide images (WSIs). At its core, HistGen employs a novel local-global hierarchical encoder that processes WSIs in a region-to-slide manner, capturing both fine-grained tissue details and broader contextual patterns. The framework first segments WSIs into regions, processes them using a pre-trained vision transformer backbone, and then hierarchically aggregates information across different scales through a series of attention-based modules. This hierarchical design allows the model to effectively manage the extreme size of WSIs while maintaining meaningful spatial relationships. Particularly noteworthy is HistGen's ability to learn robust and transferable representations through its region-aware approach, which helps bridge the gap between local tissue patterns and global slide-level characteristics.

RRT-MIL (Tang et al., 2024). RRT-MIL introduces a novel re-embedding paradigm for multiple instance learning in computational pathology that addresses a key limitation of existing approaches - the inability to fine-tune pre-extracted image features for specific downstream tasks. At its core is the Re-embedded Regional Transformer (R2T), which processes pathology images through a hierarchical structure that respects the natural organization of tissue at different scales. The framework first divides whole slide images into regions and processes them using a pre-trained feature extractor. Then, it employs two key components: a Regional Multi-head Self-attention module that captures fine-grained local patterns within each region, and a Cross-region Multi-head Self-attention module that models relationships between different regions. This regional approach allows RRT-MIL to efficiently handle the extremely large size of pathology images while maintaining the ability to capture both local and global tissue patterns.

LongMIL (Li et al., 2024a). LongMIL proposes a novel hybrid Transformer architecture specifically designed for whole slide image analysis. The method addresses key challenges in processing gigapixel pathology images by introducing a local-global attention mechanism that balances computational efficiency with modeling capability. By analyzing the low-rank nature of attention matrices in long sequences, LongMIL incorporates local attention patterns in lower layers while maintaining global context through selective attention in higher layers. This design not only reduces computational complexity but also improves the model's ability to handle varying image sizes and spatial relationships in histopathology slides.

F. Implementation Details

We implement Querent using PyTorch. For the feature extraction backbone, we utilize CPath pre-trained vision-language foundation model PLIP (Huang et al., 2023) to obtain 512-dimensional patch features. Each region contains 16/24/28 patches (depending on the datasets used), which provides a good balance between computational efficiency and contextual coverage. The model consists of 8 attention heads, with the hidden dimension set to 512.

For the region-level metadata networks (f_{min} and f_{max}), we use single-layer perceptrons with GELU activation. The query projection network f_q shares the same architecture. During the region importance estimation, we select the top-16 most relevant regions for each query patch, which empirically provides sufficient contextual information while maintaining computational efficiency.

The attention pooling network f_a consists of a two-layer MLP with a hidden dimension of 512 and GELU activation. Dropout with a rate of 0.1 is applied throughout the network to prevent overfitting. We train the model using the AdamW optimizer with a learning rate of $1e^{-4}$ for classification tasks and $2e^{-4}$ for survival analysis, with a weight decay of $1e^{-5}$. The model is trained for 50 epochs with a batch size of 1 WSI per GPU. We employ gradient clipping with a maximum norm of 1.0 to ensure stable training.

G. Additional Experiment Details

G.1. Ablation on Region-level Metadata Summarization Strategy

To thoroughly evaluate the effectiveness of our proposed min-max region-level metadata summarization approach, we conducted a comprehensive ablation study comparing it against several alternative summarization strategies:

- **Min-only**: Using only the element-wise minimum values across patches within each region
- **Max-only**: Using only the element-wise maximum values across patches within each region
- **Mean**: Using the average feature values across patches within each region
- **Mean-std**: Using both mean and standard deviation of features across patches

The evaluation framework compares the pairwise distance relationships between regions before and after summarization. For each region containing N patches with D-dimensional features ($N \times D$ matrix), we compute two distance matrices: one using the original high-dimensional features, and another using the summarized representations (e.g., D-dimensional for min/max/mean, $2D$-dimensional for min-max/mean-std). For each strategy, we computed pairwise distance matrices between regions using both the original patch features and the summarized metadata, then measured the correlation between these matrices using two metrics: Pearson correlation coefficient and Mean Squared Error (MSE), indicating how well each summarization method preserves the original structural relationships between regions.

The experiment was conducted on the TCGA-LUAD TP53 dataset, with statistical significance assessed using paired t-tests ($p < 0.005$). Our proposed min-max approach demonstrated superior performance in both metrics, achieving significantly higher correlation and lower MSE compared to all alternative strategies. This suggests that capturing both the lower and upper bounds of feature distributions within each region provides a more comprehensive and accurate representation of the region's characteristics.

Particularly noteworthy is the substantial improvement over the mean-based approaches, indicating that extreme values (minimums and maximums) carry important discriminative information that might be lost when only considering average

statistics. This aligns with pathological intuition, where both the presence of certain distinctive features (captured by maximums) and their absence (captured by minimums) can be diagnostically relevant.

G.2. Ablation on Region Importance Estimation Strategy

To evaluate the effectiveness of our proposed region importance estimation module, we conducted a detailed ablation study comparing it against two baseline approaches: random region selection and a trainable estimation side network. This study aims to validate the benefits of our query-aware dynamic region importance estimation strategy.

- **The random region selection baseline** serves as a lower bound, employing uniform random sampling to select regions for attention computation. This approach uses a fixed random seed for reproducibility and maintains the same number of selected regions as our proposed method. While simple, it helps quantify the importance of adaptive region selection.
- **The estimation side network baseline** represents a more sophisticated approach, implementing a trainable neural network that directly predicts region importance scores. This network consists of two fully-connected layers with GELU activation and processes each region independently. Unlike our proposed method, it doesn't consider query-region relationships when estimating importance.

We evaluated these approaches on both TCGA-LUAD TP53 mutation prediction and UBC-OCEAN cancer subtyping tasks. The results demonstrate that while both baseline approaches achieve reasonable performance, our query-aware estimation strategy consistently outperforms them. The performance gap is particularly pronounced in the more complex UBC-OCEAN dataset, where accurate region selection becomes crucial due to the heterogeneous nature of ovarian cancer subtypes.

For training stability, we implemented an adaptive update mechanism for the estimation module with a moving average of attention accuracy and periodic updates every 200 forward passes. This strategy helps prevent overfitting and ensures stable convergence of the importance estimation. Additionally, we employed a hybrid loss function combining binary classification and ranking components to guide the learning process effectively.

H. Complexity Analysis

We analyze the computational and memory complexity of the Query-Aware Transformer MIL method, comparing it to standard transformer self-attention. Let N denote the total number of patches in a whole slide image (WSI), d represent the hidden dimension, R be the number of regions (pages), and k indicate the number of selected regions per query. The page size p is the number of patches per region.

H.1. Standard Transformer Self-Attention

Standard transformer self-attention has quadratic computational complexity:

$$\mathcal{O}(N^2 d) \tag{13}$$

This arises from computing attention scores and weighted aggregation for all patch pairs. Memory requirements scale similarly:

$$\mathcal{O}(N^2 + Nd) \tag{14}$$

to store the full attention matrix and key/value representations. This quadratic scaling makes standard self-attention computationally prohibitive for gigapixel WSIs.

H.2. Querent: Query-Aware Dynamic Long Sequence Modeling Framework

Our method operates in three phases, with each phase optimized through efficient chunking strategies:

H.2.1. Region Metadata Computation

This phase computes min/max metadata for each region:

$$\mathcal{O}(Nd) \tag{15}$$

Memory requirements are:

$$\mathcal{O}(Rd) \tag{16}$$

where $R = \lceil N/p \rceil$ (number of regions). For constant p, R scales linearly with N. This phase is highly efficient as it requires only a single pass over the data with constant memory overhead per region.

H.2.2. Region Importance Estimation

This phase estimates relevance scores between queries and region metadata:

$$\mathcal{O}(NRd) \tag{17}$$

While the theoretical memory requirements are:

$$\mathcal{O}(NR) \tag{18}$$

in practice, we employ a chunking strategy with fixed chunk size C (typically 8192):

$$\mathcal{O}(\min(N, C) \cdot R) \tag{19}$$

This chunked processing effectively bounds the memory usage while maintaining computational efficiency through vectorized operations.

H.2.3. Selective Attention Computation

This phase performs attention only on selected regions:

$$\mathcal{O}(Nkpd) \tag{20}$$

The theoretical memory requirements are:

$$\mathcal{O}(Nkp + Nd) \tag{21}$$

However, with chunked processing (chunk size C):

$$\mathcal{O}(\min(N, C)kp + Nd) \tag{22}$$

Where k and p are constants (typically $k = 16$, $p = 16$), ensuring linear scaling with N in both computation and memory.

H.3. Total Complexity

Combining all phases, the total computational complexity is:

$$\mathcal{O}(Nd + NRd + Nkpd) \tag{23}$$

For constant k and p, this simplifies to:

$$\mathcal{O}(\mathbf{Nd} + \mathbf{NRd}) \tag{24}$$

The theoretical memory complexity is:

$$\mathcal{O}(Rd + NR + Nkp + Nd) \tag{25}$$

which simplifies to:

$$\mathcal{O}(\mathbf{NR} + \mathbf{Nd}) \tag{26}$$

However, with chunked processing, the practical memory complexity becomes:

$$\mathcal{O}(\mathbf{Rd} + \min(\mathbf{N}, \mathbf{C})\mathbf{R} + \min(\mathbf{N}, \mathbf{C})\mathbf{kp} + \mathbf{Nd}) \tag{27}$$

where C is the chunk size.

H.4. Practical Efficiency

In practice, our method achieves near-linear scaling ($\mathcal{O}(Nd)$) due to:

- Constant k and p reducing the impact of $Nkpd$
- Chunked processing with size 8192 providing bounded memory usage
- Efficient vectorized operations in PyTorch minimizing constant factors
- Memory-efficient implementations of matrix operations

Empirical results (Fig. 6) demonstrate that our method requires $\sim 1\%$ of the memory and $\sim 5\%$ of the computational cost of standard self-attention for large WSIs (100k+ patches). This significant reduction in resource requirements enables the practical processing of gigapixel-scale images while maintaining the modeling power of attention mechanisms.