EVERY SCALE HAS ITS GEOMETRY: RETHINKING POSITIONAL ENCODING FOR MULTI-RESOLUTION MULTIPLE INSTANCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple Instance Learning (MIL) enables weakly supervised analysis of Whole Slide Images (WSIs) without costly patch-level annotations, but its patch-based formulation inherently disrupts the spatial organization of tissue structures. Positional encodings (PEs) have therefore been introduced to restore spatial information lost during patchification. More recently, Multi-Resolution MIL (MR-MIL) has leveraged complementary information across magnifications by incorporating patches at multiple resolutions, better reflecting the multi-scale visual assessment performed in clinical diagnosis. However, existing PEs are not well suited to the geometry of multi-resolution WSIs (MR-WSIs): they either model spatial relationships independently within each resolution or inadequately account for the heterogeneous spatial scales induced by different magnifications. In particular, topology-aware PEs typically treat cross-resolution spatial relationships uniformly, even though the spatial domain represented by patches expands quadratically with increasing magnification. To address this limitation, we propose **H**yperbolic **H**ierarchical **P**ositional **E**ncoding (**H2PE**), a positional representation that jointly captures spatial topology and resolution hierarchy within the hyperbolic upper half-space. By leveraging the exponential capacity of hyperbolic geometry, H2PE naturally represents the hierarchical, scale-dependent structure of multi-resolution patches while preserving their spatial relationships across magnifications. H2PE is model-agnostic and can be incorporated into arbitrary MRMIL aggregation frameworks without modifying their underlying aggregation architectures. Extensive experiments on three public benchmarks demonstrate consistent improvements over existing positional encodings across multiple MRMIL frameworks. Further geometric analyses and visualizations validate that H2PE effectively captures the hierarchical organization and cross-resolution spatial structure of MR-WSIs, while providing interpretable positional representations.

1 INTRODUCTION

In the post-COVID-19 era, a surge in demand for pathological diagnoses has placed an unsustainable burden on an already limited workforce of pathologists (Bray et al., 2021; Bychkov & Schubert, 2023). While AI-driven methods offer a promising avenue to alleviate this clinical pressure, obtaining detailed annotations for gigapixel Whole Slide Images (WSIs) introduces a substantial annotation overhead for clinicians (Qu et al., 2022). To bypass this bottleneck, the deep learning (DL) community has increasingly focused on Multiple Instance Learning (MIL), a weakly supervised paradigm that classifies WSIs using only slide-level labels (Song et al., 2023).

Previous MILs have focused predominantly on developing models based on single-resolution WSIs (Ilse et al., 2018; Zhang et al., 2022). Recent advances have shifted focus to multi-resolution MIL (MRMIL), mirroring the clinical workflow where pathologists integrate multi-scale optical information (Brunyé et al., 2017). MRMIL models have demonstrated enhanced performance by exploiting both fine-grained cellular details and broader topological context (Zhang et al., 2025).MR-MIL leverages the rich information of MR-WSIs through aggregation architectures designed for multi-resolution input, spanning Nested Pyramid (Thandiackal et al., 2022; Chen et al., 2022; Xiong et al., 2023), Graph Neural Networks (GNNs) (Hou et al., 2022; Bontempo et al., 2023), and Whole-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

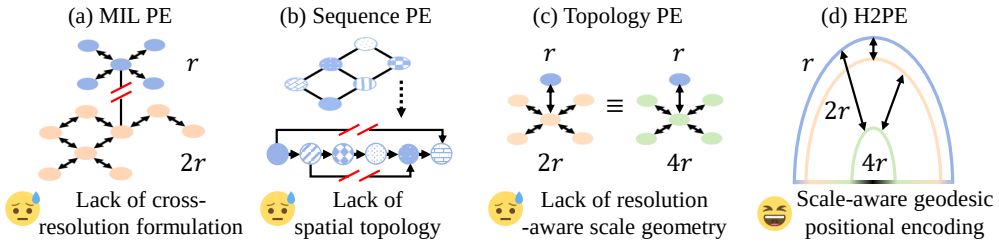


Figure 1: Comparison among PEs applied to MRMIL, where r indicates zoom level. (a) Existing MIL PEs treat relationships between resolutions as discrete, failing to explicitly capture zoomed regions corresponding to the same spatial area. (b) Sequential PEs assign a serial order to the 2D patch grid, thereby failing to preserve spatial structure. (c) Topological PEs emphasize connectivity among neighboring patches but lack awareness of the scale geometry intrinsic to resolutions. (d) H2PE formulates both resolution levels and patch relationships simultaneously within a hyperbolic space that expands exponentially.

sequence (Li et al., 2021), to construct instance bags toward visually structured representations with spatial correlations (Gadermayr & Tschuchnig, 2024; Chen et al., 2025).

Meanwhile, due to the gigapixel scale of WSIs, MIL operates on bags of instances[1] formed by extracted features from patchified tissue regions (Campanella et al., 2019). Although whole images inherently possess visually coherent spatial topology (Liu et al., 2021), this contextual structure is disrupted during patchification (Saltz et al., 2018). To compensate for this loss and preserve spatial information, positional encoding (PE) has been integrated into MIL frameworks. TransMIL (Shao et al., 2021) leverages stacked convolutions with varying kernel sizes to implicitly capture relative positional cues (Islam et al., 2020). Several works introduced a parameter-free 2D rotary positional embedding adapted for MIL (Chen et al., 2022; Nitta & Kawai, 2025). Beyond coordinate-based encodings, recent efforts have highlighted that tissue structures exhibit rotational invariance, emphasizing the local neighborhood context over rigid directional coordinates (Hong et al., 2026).

Nevertheless, designing a PE tailored for MRMIL remains an unexplored venue. The patch space in MR-WSIs exhibits unique geometric properties, which is not merely absolute, but relative (i.e., topological) and hierarchical across resolutions, spanning a Euclidean space that expands quadratically with magnification. As depicted in Fig. 1, despite these intricacies, the existing PE for MIL approaches are limited to single-resolution settings (Shao et al., 2021; Lafarge et al., 2021) and fail to formulate relationships across hierarchical MRs (Chen et al., 2022; Hong et al., 2026). Adapting general topological PEs also proves insufficient, as they fail to represent the specialized inter-resolution dynamics, most notably the quadratic spatial expansion and non-linear magnification dependencies (Shiv & Quirk, 2019; Verma et al., 2025). Practically, they treat $(r, 2r)$ and $(2r, 4r)$ patch pairs as topologically equivalent if they share an identical mapping ratio, although they represent distinctly different spatial features to clinical oracles (Kim et al., 2020).

To overcome these critical bottlenecks, we introduce **H**yperbolic **H**ierarchical **P**ositional **E**ncoding (**H2PE**), a novel positional formulation specifically tailored to the unique topological geometry of MR-WSI. As diagnostic features in WSIs are invariant to affine transformations (Lafarge et al., 2021; Kang et al., 2023), irregular tissue patterns strictly depend on relative topological arrangements (Doyle et al., 2008). Simultaneously, patches across resolutions form non-Euclidean parent-child hierarchies where space expands quadratically, as the resolution increases. Built upon the hyperbolic upper half-space model, H2PE seamlessly captures this exponential spatial expansion while explicitly modeling parent-child and sibling relationships. Unlike conventional topological PE that treats cross-resolution distances homogeneously, H2PE preserves resolution heterogeneity within a unified Riemannian manifold. We design H2PE to integrate with arbitrary MRMIL aggregation mechanisms. Extensive experiments conducted across three challenging public benchmarks demonstrate that H2PE yields consistent performance gains over competing PEs, proving its efficacy for MR-WSIs that scale quadratically. Further analytical experiments elucidate the fundamental principles underlying H2PE. Visualization of PEs reveals the limitations of existing approaches on MR-WSIs while demonstrating that H2PE improves model interpretability.

[1]For clarity, we use ”patch” and ”instance” interchangeably.

2 Related Work

2.1 Multiple Instance Learning for Digital Pathology

DL based WSI analysis is inherently framed as a weakly supervised MIL due to the gigapixel resolution and the scarcity of pixel-level annotations (Gadermayr & Tschuchnig, 2024). Early approaches relied on pooling operators such as mean, max, and min, which often suffer from information loss (Ramon et al., 2000). AB-MIL (Ilse et al., 2018) and TransMIL (Shao et al., 2021) introduced strong performance and interpretability based on gate- and self-attention, respectively. Attention mechanism have since become the foundation of MIL, being adopted across diverse strategies such as dual-tier frameworks (Li et al., 2021), pseudo-bag (Zhang et al., 2022), clustering (Lu et al., 2021), and sequential formulations (Yang et al., 2024). More recently, the community has been broadening the scope by transfer learning (Shao et al., 2025) and fusing prior knowledge from large language models (LLMs) (Liu et al., 2025).

In clinical practice, pathologists routinely switch between low and high magnifications to integrate macroscopic tissue context (Brunyé et al., 2017). Emulating this diagnostic workflow, Multi-Resolution MIL (MRMIL) has recently emerged as the prevailing paradigm (Zhang et al., 2025). MRMIL frameworks generally fall into three principal categories based on their hierarchical modeling strategy: Nested Pyramid (e.g., hierarchical zoom) paradigms (Thandiackal et al., 2022; Chen et al., 2022; Xiong et al., 2023), which model the direct containment relationships between fine-grained patches and their corresponding broad contextual surroundings; hierarchical GNN-based structures, which represent MR topological (Hou et al., 2022; Bontempo et al., 2023) and hyperbolic (Huang et al., 2025) patch graphs to capture both spatial proximities and cross-scale contextual interactions; and Whole-sequence (Li et al., 2021), which extract multi-scale patches at predefined magnification levels and fuse multi-scale features.

2.2 Positional Encoding for Multiple Instance Learning

PE has been far more extensively investigated in language and generalized domains (Dufter et al., 2022). Absolute schemes add a sinusoidal (Vaswani et al., 2017) or learned vector to each token (Shaw et al., 2018). Relative schemes make the attention logit a function of the offset, as a trainable bias (Raffel et al., 2020), a rotation of queries and keys (Su et al., 2024), and a distance penalty (Press et al., 2021). Vision models extend this to a 2D lattice (Liu et al., 2021; Chu et al., 2021). Topological tree encodings compose affine maps along the path from a root (Shiv & Quirk, 2019; Cantürk et al., 2023), where graph encodings derive coordinates from Laplacian eigenvectors (Dwivedi et al., 2021) and homology (Verma et al., 2025). Hyperbolic positional encodings have been proposed for 1D language sequences (Dai et al., 2025; Angelotti, 2023).

In MIL for pathology, most aggregators treat bags as permutation-invariant, collapsing the spatial topology of instances, i.e., original WSI (Chen et al., 2025). TransMIL (Shao et al., 2021) is the first to highlight the importance of relative spatial locations and inter-patch relations, proposing a 2D convolution-based PE. Subsequently, RoTransMIL (Nitta & Kawai, 2025) incorporated the rotation-invariant characteristics inherent to WSIs. Although there have been occasional attempts to employ PE in MR-WSI (Chen et al., 2022; Hong et al., 2026), they have remained limited to modeling relations discretely within individual resolutions.

3 Method

3.1 Preliminaries

MIL Properties. Let $\mathcal{D} = \{(W_a, L_a)\}_{a=1}^{N_\mathcal{D}}$ denote a dataset comprising $N_\mathcal{D}$ WSIs along with their corresponding slide-level labels. Within the MIL paradigm, each WSI W_a is partitioned into N_a non-overlapping patches, represented as $\{w_{(a,i)}\}_{i=1}^{N_a}$. Each patch $w_i := w_{(\cdot,i)}$ is designated by its unique patch index coordinates (x_i, y_i). A pre-trained feature extractor projects these patches into a bag of instance features $\mathbf{V}_a = \{v_{(a,i)}\}_{i=1}^{N_a}$, where $v_i := v_{(\cdot,i)} \in \mathbb{R}^d$ and d denotes the feature dimension. The (MR)MIL aggregator $f_\theta(\cdot)$ generates the slide-level prediction $\hat{L}_a = f_\theta(\mathbf{V}_a)$.

Multi-Resolution and Hyperbolic Upper Half-Space Model. In a MR framework, W_a contains

tissue images at or below the maximum scan resolution R, i.e., $W_a = \{W_a^r | r \leq R\}$. Since digital slide scanners adopt a differing magnification policy according to manufacturers, Microns per Pixel (MPP) is universally employed as a standardized resolution metric (Brixtel et al., 2022). An image in mMPP indicates that each digital pixel corresponds to $m\,\mu m$ of real tissue, where lower values of m represent higher resolutions (i.e., $m \propto 1/r$). Consequently, shifting from mMPP to m/rMPP increases the resolution by r times and scales the total patch count by r^2. This shows the quadratic growth of the patch space as one zooms into finer details. Embedding such exponentially branching structures into a polynomialy scaling Euclidean space inherently induces metric distortion (Sala et al., 2018). In contrast, hyperbolic space, characterized by exponential volume growth, can naturally accommodate exponential tree structures without distortion (Fig. 2). Therefore, we formulate the spatial positioning of MR-WSI within the upper half-space (UHS) model, $\mathbb{H}^3 = \{(x, y, z) \in \mathbb{R}^3 : z > 0\}$, where z is a resolution component. In a MR-WSI, zooming intrinsically corresponds to transformations, which precisely match the isometry group of the UHS:

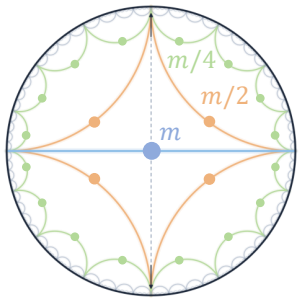


Figure 2: The MR-WSI quadtree as a geodesic lamination in the Poincaré disk. Halving the MPP yields child geodesics while zooming in advances toward the boundary (i.e., $m \to 0$), where hyperbolic space expands exponentially.

$$ds^2 = \frac{dx^2 + dy^2 + dz^2}{z^2}, \tag{1}$$

where scaling by r preserves distance (i.e., $ds^2 = \frac{r^2(dx^2+dy^2+dz^2)}{r^2 z^2}$). By leveraging the UHS model, we naturally incorporate the physical symmetries inherent to the problem.

3.2 H2PE: Hierarchical Hyperbolic Positional Encoding

3.2.1 Spatial Representation of Patches

Since (x_i, y_i) of the patch w_i, which is extracted from m_iMPP, are level-specific grid indices with varying physical scales (i.e., $(x_\cdot, y_\cdot) = (x_i, y_i)$ at m_iMPP and $m_{j\neq i}$MPP correspond to different tissue locations), we convert all coordinates across levels into a single common unit, where P is pixel size of the patch width (height):

$$z_i = m_i \times P, \qquad c_i = z_i \times \left(x_i + \tfrac{1}{2}, y_i + \tfrac{1}{2}\right) \in \mathbb{R}^2, \tag{2}$$

Here, z_i denotes the actual physical tissue size covered by w_i in micrometers, and c_i represents its physical center coordinates. We aim to construct a positional representation using $z_\cdot$ and $c_\cdot$, which can be seamlessly integrated into arbitrary MRMIL architectures.

3.2.2 Structural Distance

The unique spatial representation of patch w_i, mapped onto the UHS coordinates $(z_i, c_i) \in \mathbb{R}^3$ through Equ. 2, defines the hyperbolic distance to another patch w_j in a closed-form:

$$d_{ij} = \operatorname{arcosh}\left(1 + \frac{\|c_i - c_j\|^2 + P^2(m_i - m_j)^2}{2 \times P^2 m_i m_j}\right) \in \mathbb{R}. \tag{3}$$

The metric d_{ij} reflects the physical rationale of MR-WSI modeled in hyperbolic UHS. At the same resolution (i.e., $m_i = m_j$), d_{ij} depends solely on the magnitude $\|c_i - c_j\|^2$, rendering it invariant to arbitrary rotations, which is a property that naturally mirrors the pathological fact that tissue sections lack a canonical orientation (Kim et al., 2020). Furthermore, for two patches at different resolutions covering the exact same tissue region, d_{ij} scales with the MPP ratio $\delta = m_i/m_j$ as $\operatorname{arccosh}((\delta + \delta^{-1})/2) \approx \ln \delta$. Consequently, the distance accumulates linearly with each level shift. This captures the proportional scale relationships that conventional topological PEs, which merely count hop distances, fail to represent. We provide detailed proofs in the Appendix.

3.2.3 Asymmetric Directionality

While d_{ij} is symmetric by definition (i.e., $d_{ij} = d_{ji}$), top-down zooming observation (i.e., parent to child) and bottom-up zooming out observation (i.e., child to parent) hold distinct epistemological statuses in clinical pathology. For a child patch, the parent is a uniquely determined entity,

whereas for a parent patch, a child represents an ambiguous target that must be selected among r^2 candidates. Furthermore, pathognomonic findings such as mitotic figures or micrometastases are physically confined to small scales and cannot be reliably identified under low power observation alone (Ehteshami Bejnordi et al., 2017), demonstrating that the relationship between low and high magnifications is naturally asymmetric. We explicitly model this asymmetry not through the inherently symmetric metric, but by introducing separate directional terms:

$$g_i = \ln m_i, \quad \Delta_{ij} = \ln \frac{m_i}{m_j}, \quad \gamma_{ij} = \gamma_0 \cdot \left(\frac{m_{\text{ref}}}{\min(m_i, m_j)} \right)^{\beta}, \quad s_{i \to j} = d_{ij} + \gamma_{ij} \Delta_{ij}. \tag{4}$$

Δ_{ij} possesses complete antisymmetry, where $\Delta_{ij} = -\Delta_{ji}$, acting as a function purely of the ratio between two MPPs without depending on P. The γ_{ij} is designed to be exponentially proportional to how much finer the higher resolution of the pair, $\min(m_i, m_j)$, is relative to the coarsest MPP m_{ref} in the WSI. Here, $\gamma_0 > 0$ is a learnable scale parameter setting the base magnitude of γ_{ij} when comparing the coarsest resolutions with $\min(m_i, m_j) = m_{\text{ref}}$, while $\beta \geq 0$ is a learnable exponent controlling its sensitivity to resolution differences. This consistently extends the design principle of exponential scale response adopted in Section 3.1 to the directional term. When $m_i = m_j$, Δ_{ij} evaluates to zero, allowing $s_{i \to j}$ to naturally reduce to the symmetric d_{ij}, which is fully aligned with the requirement that directional asymmetry is unnecessary within the same resolution (MPP).

3.2.4 Integration Scheme for Arbitrary MRMIL Architectures

Arbitrary MRMIL aggregators $f_\theta(\cdot)$, consistently comprise two entry points: one that receives patch features v_i and another that computes interaction scores A_{ij} between patch pairs. We inject g_i into the former and $s_{i \to j}$, which is regarded as an indicator of directional design, into the latter. For actual injection, we explicitly feed $s_{i \to j}$ to decompose d_{ij} and $\gamma_{ij}\Delta_{ij}$ by component without loss of information. Instead of adding these raw scalars to v_i and A_{ij} as fixed constant multiples, we project them through multi-layer perceptrons (MLP), allowing the network to learn the integration mechanism directly from data:

$$e_i = \text{MLP}_g\big(g_i\big) \in \mathbb{R}^d, \qquad b_{ij} = \text{MLP}_r\big(\text{Concat}(d_{ij}, \gamma_{ij}\Delta_{ij})\big) \in \mathbb{R}. \tag{5}$$

The resolution embedding e_i is added to the patch feature to yield $v_i' = v_i + e_i$. Since most MRMIL architectures rely on Transformers (Vaswani et al., 2017) and GNNs, we design specialized injection methods for these frameworks alongside a versatile option for other structural designs.

Transformers. As self attention already computes $A_{ij} = q_i \cdot k_j / \sqrt{d_k}$ for every pair (i, j), $A'_{ij} = A_{ij} - b_{ij}$ is applied right before softmax.

GNNs. Since the message passing mechanism is formulated on the underlying edge set, $A'_{ij} = A_{ij} - b_{ij}$ is identically incorporated into the edge weight or interaction coefficient term A_{ij}.

Etcetera. For architectures that do not expose pairwise terms A_{ij}, we append a minimal relational residual module receiving only b_{ij} as input, independently of the backbone. The set of patch pairs E is defined by same level k nearest neighbor relations and quadtree parent child relationships. Since d_{ij} is additive along quadtree paths, performing $\mathbb{L}$ layers of message passing over E propagates relationships between any two patches with a quadtree path length up to $\mathbb{L}$. We define a relational residual module that receives only $\{b_{ij}\}_{(i,j)\in E}$ as input:

$$h_i = \sum_{j:(i,j)\in E} \text{softmax}_{j:(i,j)\in E}(-b_{ij}) \cdot W v'_j \in \mathbb{R}^d, \qquad v''_i = v'_i + h_i, \tag{6}$$

where $W \in \mathbb{R}^{d \times d}$ denotes a learnable linear projection. Since this scheme makes no assumptions about backbone, it serves as a minimal baseline applicable to any arbitrary MRMIL architecture. We present the detailed MLP recipe in the Appendix.

4 Experiment

4.1 Experiment Settings

Dataset. The CAMELYON16 dataset (Ehteshami Bejnordi et al., 2017) comprises 400 hematoxylin and eosin (H&E) stained sentinel lymph node WSIs from breast cancer patients, annotated with binary labels for lymphatic metastasis diagnosis. The BRACS dataset Brancati et al. (2022) comprises

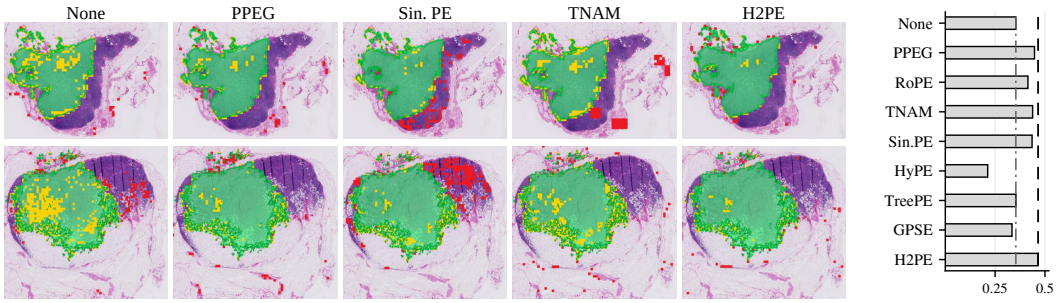
Table 1: Quantitative performance results. **Bold** text indicates the best performance, and underlined text highlights the second best.

MR-MIL	PE	CAMELYON16		BRACS-C		BRACS-F		BCNB	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
ZoomMIL	-	86.1$_{1.63}$	91.0$_{2.37}$	66.7$_{2.18}$	85.7$_{1.18}$	43.7$_{3.17}$	80.8$_{0.53}$	73.7$_{1.46}$	77.9$_{2.10}$
	PPEG	88.8$_{1.52}$	91.9$_{1.50}$	66.0$_{1.17}$	82.7$_{1.25}$	43.7$_{2.06}$	80.3$_{0.85}$	72.6$_{2.28}$	77.5$_{1.79}$
	RoPE	89.1$_{2.86}$	92.5$_{1.90}$	69.9$_{1.97}$	88.0$_{1.61}$	44.1$_{2.25}$	81.0$_{1.17}$	74.1$_{2.53}$	78.3$_{2.07}$
	TNAM	88.7$_{0.62}$	91.2$_{2.15}$	69.2$_{1.97}$	86.5$_{2.30}$	46.7$_{2.01}$	81.3$_{1.13}$	73.6$_{2.71}$	77.2$_{2.08}$
	Sin.PE	88.8$_{2.80}$	91.5$_{3.82}$	68.5$_{2.87}$	87.0$_{1.03}$	45.3$_{2.48}$	79.7$_{1.36}$	73.3$_{2.17}$	76.4$_{2.48}$
	HyPE	86.7$_{2.16}$	87.2$_{2.37}$	68.0$_{2.10}$	86.6$_{1.57}$	44.1$_{3.45}$	79.2$_{1.22}$	71.5$_{0.81}$	77.7$_{0.64}$
	TreePE	88.5$_{0.58}$	90.3$_{2.25}$	65.1$_{2.01}$	83.3$_{0.97}$	43.5$_{3.36}$	81.5$_{0.66}$	74.0$_{1.13}$	78.2$_{2.37}$
	GPSE	88.5$_{2.88}$	92.0$_{3.38}$	69.2$_{1.52}$	87.8$_{1.37}$	47.1$_{2.81}$	81.4$_{0.45}$	74.3$_{1.93}$	78.4$_{1.71}$
	H2PE	**90.5**$_{1.42}$	**93.5**$_{2.11}$	**71.5**$_{2.66}$	**88.1**$_{0.53}$	**50.6**$_{2.62}$	**82.0**$_{0.42}$	**74.4**$_{0.92}$	**79.5**$_{1.34}$
H^{2}MIL	-	86.4$_{1.16}$	86.4$_{0.52}$	67.8$_{1.03}$	85.7$_{0.80}$	45.5$_{3.30}$	82.2$_{0.39}$	70.1$_{1.38}$	73.6$_{0.87}$
	PPEG	88.5$_{0.31}$	**91.4**$_{0.70}$	67.6$_{1.52}$	85.5$_{0.60}$	46.0$_{2.72}$	78.8$_{0.61}$	72.2$_{1.87}$	75.9$_{1.58}$
	RoPE	87.1$_{1.44}$	87.6$_{0.94}$	67.1$_{1.38}$	88.0$_{0.87}$	47.4$_{0.86}$	81.8$_{0.82}$	68.4$_{0.76}$	72.1$_{1.49}$
	TNAM	89.0$_{1.65}$	88.1$_{1.00}$	67.4$_{2.36}$	87.8$_{0.83}$	47.6$_{1.56}$	81.1$_{0.55}$	69.7$_{1.89}$	73.3$_{1.75}$
	Sin.PE	87.0$_{1.33}$	87.5$_{1.07}$	67.4$_{1.17}$	84.9$_{0.63}$	45.8$_{2.23}$	82.4$_{0.54}$	69.0$_{1.76}$	73.2$_{1.48}$
	HyPE	83.6$_{1.14}$	85.7$_{0.75}$	67.1$_{1.17}$	88.0$_{0.78}$	43.0$_{2.37}$	80.1$_{0.55}$	71.0$_{1.09}$	72.9$_{0.91}$
	TreePE	87.6$_{1.20}$	89.2$_{0.58}$	67.4$_{2.37}$	87.8$_{0.72}$	46.0$_{1.78}$	80.8$_{0.27}$	70.7$_{1.06}$	72.1$_{1.66}$
	GPSE	86.4$_{1.16}$	87.2$_{1.26}$	68.3$_{4.40}$	86.7$_{2.69}$	46.2$_{2.66}$	82.1$_{1.05}$	70.2$_{2.58}$	71.9$_{4.61}$
	H2PE	**90.4**$_{1.35}$	91.1$_{1.59}$	**69.4**$_{2.00}$	**88.2**$_{1.19}$	**48.5**$_{2.10}$	**82.7**$_{0.66}$	**73.7**$_{0.42}$	**77.0**$_{0.70}$
DSMIL	-	86.7$_{9.82}$	91.5$_{6.17}$	65.5$_{3.78}$	83.6$_{2.30}$	44.6$_{2.11}$	76.7$_{1.52}$	65.5$_{2.03}$	68.6$_{1.46}$
	PPEG	90.2$_{1.35}$	92.9$_{1.88}$	67.8$_{3.85}$	**85.5**$_{2.54}$	45.5$_{2.78}$	78.5$_{1.80}$	**71.5**$_{2.38}$	**77.1**$_{3.45}$
	RoPE	90.2$_{0.79}$	92.1$_{2.22}$	62.5$_{2.68}$	79.9$_{3.82}$	44.6$_{3.03}$	76.8$_{2.20}$	67.1$_{2.45}$	70.2$_{5.20}$
	TNAM	88.4$_{1.62}$	90.8$_{2.45}$	66.9$_{3.03}$	83.7$_{2.22}$	44.4$_{3.13}$	79.5$_{2.05}$	68.5$_{2.09}$	71.5$_{2.21}$
	Sin.PE	90.8$_{1.03}$	93.4$_{2.12}$	68.0$_{2.45}$	83.6$_{2.33}$	43.5$_{2.23}$	77.7$_{1.67}$	63.3$_{1.79}$	67.5$_{1.76}$
	HyPE	86.7$_{4.79}$	86.2$_{5.87}$	68.5$_{0.92}$	84.2$_{1.61}$	44.1$_{2.78}$	78.1$_{2.12}$	66.8$_{4.17}$	69.9$_{4.31}$
	TreePE	**91.2**$_{2.33}$	93.4$_{0.96}$	68.0$_{1.69}$	84.0$_{1.83}$	45.3$_{4.40}$	77.2$_{2.82}$	66.5$_{2.66}$	70.2$_{2.78}$
	GPSE	90.4$_{1.26}$	92.5$_{1.62}$	68.0$_{1.98}$	84.0$_{1.62}$	45.1$_{3.80}$	77.6$_{1.96}$	70.7$_{2.27}$	74.9$_{1.83}$
	H2PE	**91.2**$_{1.16}$	**94.4**$_{1.80}$	**69.4**$_{2.58}$	85.1$_{1.85}$	**48.3**$_{1.45}$	**80.0**$_{1.90}$	**71.5**$_{2.51}$	75.0$_{2.30}$

547 H&E stained breast carcinoma WSIs hierarchically categorized into 3 coarse and 7 fine classes: Benign (Normal, Pathological Benign, Usual Ductal Hyperplasia), Atypical (Flat Epithelial Atypia, Atypical Ductal Hyperplasia), and Malignant (Ductal Carcinoma in Situ, Invasive Carcinoma). We evaluate performance on both BRACS-Coarse (C) and -Fine (F) setups. The BCNB dataset (Xu et al., 2021) comprises H&E stained WSIs of early-stage breast cancer core-needle biopsies, accompanied by clinical annotations of axillary lymph node (ALN) metastasis status. Following the original setup, we evaluate a binary ALN classification task that distinguishes node-negative from node-positive cases. For these datasets, tissue regions were first segmented using Otsu's thresholding algorithm (Otsu, 1979). Patches of size $P = 256$ pixels were extracted at 2, 1, and 0.5 MPPs, after which their feature representations were generated using a pretrained encoder (Lu et al., 2024).

MRMIL Architectures. To assess the adaptability of PEs to arbitrary MRMIL aggregation schemes, our experiments incorporate well established baseline architectures thoroughly validated and widely recognized in each aggregation category. (i) Nested Pyramid: ZoomMIL (Thandiackal et al., 2022) incorporates an end to end differentiable zooming mechanism that selectively aggregates regions across magnifications. (ii) GNNs: H^{2}MIL (Hou et al., 2022) learns hierarchical representations via resolution aware attention convolutions on a heterogeneous graph. (iii) Whole-sequence: DSMIL (Li et al., 2021) is a dual-stream framework that integrates MR instances using distance based attention relative to max pooled critical instances.

Comparison Methods. We include both PEs tailored specifically for WSIs and competitive general-purpose approaches for comparison. PPEG (Shao et al., 2021) reshapes the patch sequence into a 2D grid and injects local spatial information through convolution kernels of varying sizes. The PE in RoTransMIL (Nitta & Kawai, 2025) applies parameter free 1D and 2D rotary PEs directly within the dot product attention mechanism (Su et al., 2024). TNAM (Hong et al., 2026) serves as a neighbor aggregation module that captures spatial information between adjacent patches using attention based mechanisms. As general approaches, Sinusoidal (Sin.) PE (Vaswani et al., 2017) serves as a base-



(a) Attention map visualizations. ■, ■, and ■ represent regions of true positives, false negatives, and false positives, respectively.

(b) mIoU on the entire test set.

Figure 3: DSMIL attention maps and mIoUs at the finest MPP for different PEs.

line method that adds fixed sinusoidal functions of 1D spatial indices directly to token embeddings. TreePE (Shiv & Quirk, 2019) encodes the path from the tree root to each node via a composition of learnable affine transformations. GPSE (Cantürk et al., 2023) is a general purpose graph based encoder pretrained to simultaneously reconstruct diverse topological PEs. HyPE (Angelotti, 2023) employs hyperbolic functions to encode relative positions of sequences as attention biases.

Implementation Details. The optimization is conducted using Adam (Kingma, 2014) with an initial learning rate of 1×10^{-4} and momentum parameters $\beta = (0.9, 0.999)$, where the learning rate is decayed to zero over the training epochs via a cosine annealing scheduler (Loshchilov & Hutter, 2016). All experiments are executed multiple times on an NVIDIA® DGX Spark™. To ensure fair evaluation, each baseline method is evaluated across multiple independent runs using its respective optimal hyperparameter configuration.

4.2 Quantitative Results

Tab. 1 reports the results of three MRMIL aggregators combined with each PE applicable to MR-WSIs. Although PEs generally improve MRMIL, every existing PE degrades at least one aggregator below its PE-free counterpart, echoing the observations of Hong et al. (2026) and underscoring that a PE must match the structural complexity of MR-WSIs. PPEG is competitive on DSMIL, whose concatenated pillars collapse the bag into a single grid, yet its per-resolution convolution falls below the PE-free baseline in five of eight entries on ZoomMIL, where zooming across resolutions is essential, and its cost on large instance sets is further examined in Section 4.3.3. Sin. PE and RoPE prioritize token discrimination over resolution relationships, so their gains largely vanish on H2MIL, whose message passing relies on explicit zooming structure. For fine subtyping, where cellular morphology and growth patterns across magnifications are critical, local neighborhood smoothing via TNAM fails to capture cross-scale cues and yields unstable results. HyPE embeds grid coordinate distances as hyperbolic biases without unifying them in a physical metric space, recording the lowest average AUC for every aggregator, while TreePE and GPSE replace continuous geometry along resolution ratios with discrete hop counts and remain competitive only sporadically. H2PE resolves these collective limitations by fully exploiting the multiscale context inherent in MR-WSIs, ranking first in 21 of 24 entries without ever falling below the PE-free baseline. Its margins are most pronounced on BRACS-F, reaching 3.5% in accuracy where multiscale morphology matters most. These consistent gains across MRMILs designed for inter-resolution modeling, together with its leading average even on DSMIL, confirm that H2PE supplies an effective spatial inductive bias even when structural assumptions are minimal.

4.3 Further Analysis

4.3.1 Diagnostic Explainability

Protocol. To verify the improvement in diagnostic interpretability afforded by PEs, we perform visual analysis using DSMIL, which enables the extraction of attention maps (see Fig. 3). The CAMELYON16 dataset, featuring oracle provided pixel level annotations, was selected for evalua-

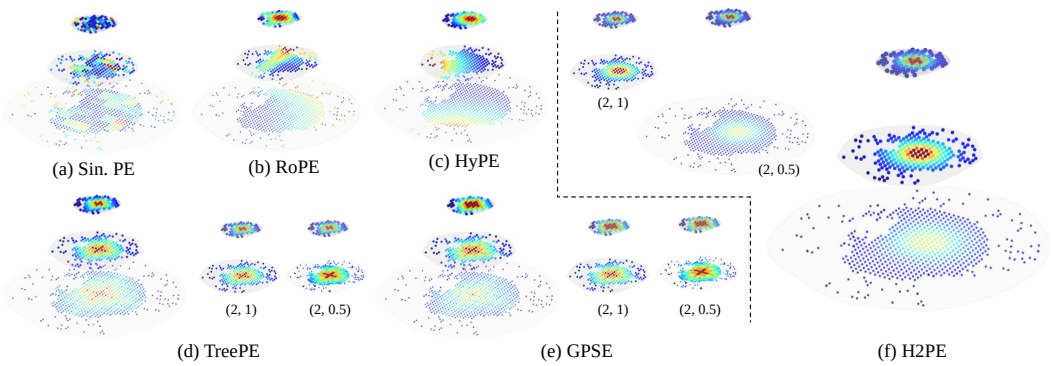


Figure 4: PE visualization across multiple MPPs, where $(\cdot, \cdot)$ denote a pair of two subsampled scales.

tion. As a quantitative metric, we evaluate the Intersection over Union (IoU) between the regions enclosed by ground truth polygons and the top k attention regions. Within WSIs at $0.5\,\text{MPP}$, any patch whose area is covered by polygons by 10% or more is designated as an annotated patch. Note that for each individual WSI, k is set equal to the total number of annotated patches.

Results. Fig. 3a displays the attention maps generated by the protocol. While the baseline successfully highlights the overall lesion area, it exhibits a high rate of false negatives ■ (FNs) . Although all PEs adequately mitigate FNs, Sin. PE and TNAM produce overestimations that increase false positive ■ (FP) regions. PPEG achieves the most noticeable enhancement in interpretability among the comparison methods, yet it still shows FPs along certain boundaries. In contrast, H2PE appropriately resolves both FNs and FPs, thereby elevating the interpretability for WSI diagnosis. Fig. 3b presents the mIoU across all test samples. A performance drop is observed for PEs that formulate topological relationships. Conversely, PPEG and TNAM, which are tailored for MIL, yield high mIoU values, demonstrating that properly defining intra resolution relationships is crucial for improving interpretability. H2PE successfully captures both inter resolution and intra resolution dynamics, leveraging these advantages to achieve superior performance.

4.3.2 Visualization

Fig. 4 visualizes H2PE alongside comparison PEs, excluding PPEG and TNAM which rely solely on learnable parameters. The structural connectivity of an actual MR-WSI is depicted top-down across hierarchy levels, where distance or vector similarity among encoding components is colorized via the jet colormap. (a) Sin. PE focuses primarily on component distinction, failing to capture spatial transitions and intrinsic visual properties because it assumes neither the expanding geometry nor the scale connectivity of MR-WSIs. (b) RoPE and (c) HyPE represent neighbor relationships within the identical resolution, yet remain completely oblivious to the inter resolution hierarchy. (d) TreePE and (e) GPSE capture both intra scale spatial neighbor relations and inter resolution topological links, as evidenced by the smooth continuity of the colormap across scales. However, for heterogeneous pairs involving different MPP combinations, they assign identical topological representations due to their reliance on one hop relational awareness without absolute resolution information. (f) H2PE addresses intra scale instance relations with inter scale spatial structures. In heterogeneous MPP settings, rather than depending solely on topology. H2PE embeds identical resolution levels into a unified hyperbolic space to encode distinct multi scale relationships, as reflected in the joint behavior of distance and MPP scales.

4.3.3 Computational Efficiency of Positional Encodings

We evaluate the computational efficiency of PEs applied to MRMIL and their data efficiency with respect to increasing MR patches using ZoomMIL on the CAMELYON16 dataset.

Compute Efficiency. Fig. 5a shows performance relative to actual runtime per WSI. All PEs spent under $1000\,\text{ms}$ (i.e., $1\,\text{second}$) on position construction. The fact that certain encodings underperformed the PE free baseline despite additional computational cost highlights the importance of incorporating appropriate geometry for MRMIL. H2PE achieves superior performance in roughly $0.01\,\text{seconds}$. A comprehensive computational complexity analysis is provided in the Appendix.

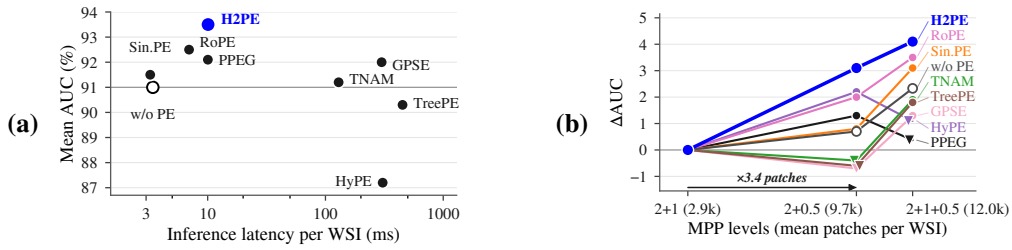


Figure 5: (a) Performance of PEs against real execution time and (b) data efficiency with information gain from MPP expansion.

Data Efficiency. Although high-resolution patches provide richer information at substantially higher data and preprocessing cost, Fig. 5b shows that $(2, 0.5)$MPP, with 3.4 times more patches than $(2, 1)$MPP, degrades the topology-aware encodings of TNAM, TreePE, and GPSE, and although they recover at $(2, 1, 0.5)$MPPs, the additional level in turn reverses the gains of PPEG and HyPE. Only Sin.PE and RoPE, which disregard cross-resolution topology, improve monotonically along with the PE-free baseline, yet their gains remain well below that of H2PE, as they merely accumulate instances rather than exploit their scale relations. H2PE is thus the only topology-aware encoding that improves monotonically with the largest gain at every budget, highlighting its ability to structure expanded spatial geometry and efficiently process scaled instances.

4.3.4 Ablation Study

We conduct an ablation study to verify the contributions of the components proposed in H2PE, specifically the UHS model based distance, the asymmetric view $s_{i\to j}$ (i.e., without Δ_{ij}), and the node level scale MLP$_g$ (Fig. 6). Replacing the UHS metric with simple Euclidean distance degrades overall performance, demonstrating that the exponential expansion of UHS space effectively captures distinct zoom level geometries in MR settings. Omitting directionality yields the most substantial performance drop, implying that MRMIL instances require explicit directional distinction along parent-child hierarchies. Furthermore, removing MLP$_g$ confirms that projecting positional digits into higher dimensional embeddings provides superior feature representations compared to simply repeating raw encodings.

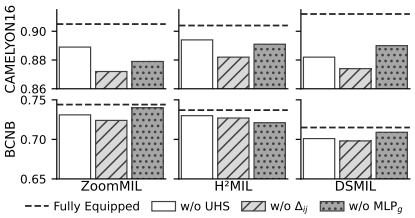


Figure 6: Ablation plots on MRMILs.

5 Conclusion

In this work, we propose Hyperbolic Hierarchical Positional Encoding (H2PE) to address the resolution-dependent spatial relationships inherent to MRMIL, which are not adequately captured by existing PEs. H2PE jointly represents spatial position and scale in the hyperbolic upper half-space, allowing the positional representation to preserve spatial topology while modeling the scale-dependent geometry across resolutions. The formulation further incorporates directional cross-resolution relationships and can be integrated into different MRMIL aggregation architectures without modifying their underlying aggregation mechanisms. Experiments on challenging public benchmarks and representative MRMIL frameworks demonstrate the effectiveness of H2PE, while analytical experiments and visualizations support its geometric properties and improved spatial interpretability. These results indicate that modeling resolution as an intrinsic component of positional geometry provides a principled way to represent the multi-scale spatial structure of MR-WSIs.

Limitations and Future Work. Although H2PE is defined for arbitrary magnifications through physical MPP values, our experiments use dyadically spaced levels and a fixed patch size. In future work, we aim to evaluate non-dyadic, scanner-specific magnification sets and deeper pyramids, and to exploit this physical coordinate system for further cross-scanner generalization.

AI USE STATEMENT

In this work, we used generative AI tools for grammar, spelling, and punctuation correction, as well as for feedback on the presentation and organization of experiments. We have reviewed all AI-assisted work and take full responsibility for the final content of this work.

ETHICS STATEMENT

This study was performed in line with the principles of the Declaration of Helsinki. This research study was conducted retrospectively using human subject data made available in open access by Ehteshami Bejnordi et al. (2017), Brancati et al. (2022) and Xu et al. (2021). Ethical approval was not required as confirmed by the license attached with the open access data.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide an anonymized repository at `https://anonymous.4open.science/r/H2PE-git-E73D/`, which contains the source code and implementation of our method. All datasets used in our experiments are publicly available. For data preprocessing and experimental protocols, we follow the standard protocols established by the respective prior works. For the MRMIL architectures evaluated in our experiments, we refer to and build upon the official implementations provided by the respective works. Detailed descriptions of the proposed method, experimental settings, and additional analyses are provided in the main paper and appendix.

REFERENCES

Giorgio Angelotti. Hype: Attention with hyperbolic biases for relative positional encoding. *arXiv preprint arXiv:2310.19676*, 2023.

Gianpaolo Bontempo, Angelo Porrello, Federico Bolelli, Simone Calderara, and Elisa Ficarra. Dasmil: Distilling across scales for mil classification of histological wsis. In *International conference on medical image computing and computer-assisted intervention*, pp. 248–258. Springer, 2023.

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022: baac093, 2022.

Freddie Bray, Mathieu Laversanne, Elisabete Weiderpass, and Isabelle Soerjomataram. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, 127 (16):3029–3030, 2021.

Romain Brixtel, Sebastien Bougleux, Olivier Lézoray, Yann Caillot, Benoit Lemoine, Mathieu Fontaine, Dalal Nebati, and Arnaud Renouf. Whole slide image quality in digital pathology: review and perspectives. *IEEE Access*, 10:131005–131035, 2022.

Tad T Brunyé, Ezgi Mercan, Donald L Weaver, and Joann G Elmore. Accuracy is in the eyes of the pathologist: the visual interpretive process and diagnostic accuracy with digital whole slide images. *Journal of biomedical informatics*, 66:171–179, 2017.

Andrey Bychkov and Michael Schubert. Constant demand, patchy supply. *Pathologist*, 88(88): 18–27, 2023.

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

Semih Cantürk, Renming Liu, Olivier Lapointe-Gagné, Vincent Létourneau, Guy Wolf, Dominique Beaini, and Ladislav Rampášek. Graph positional and structural encoder. *arXiv preprint arXiv:2307.07107*, 2023.

Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16123–16134. IEEE, 2022.

Xiwen Chen, Peijie Qiu, Wenhui Zhu, Hao Wang, Huayu Li, Xuanzhao Dong, Xiaotong Sun, Xiaobing Yu, Yalin Wang, Abolfazl Razi, et al. Cracking instance jigsaw puzzles: An alternative to multiple instance learning for whole slide image analysis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21353–21363. IEEE, 2025.

Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021.

Chang Dai, Hongyu Shan, Mingyang Song, and Di Liang. Hope: Hyperbolic rotary positional encoding for stable long-range dependency modeling in large language models. *arXiv preprint arXiv:2509.05218*, 2025.

Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *2008 5th IEEE international symposium on biomedical imaging: from nano to macro*, pp. 496–499. IEEE, 2008.

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. Position information in transformers: An overview. *Computational Linguistics*, 48(3):733–763, 2022.

Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph neural networks with learnable structural and positional representations. *arXiv preprint arXiv:2110.07875*, 2021.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Camelyon16 Consortium, Meyke Hermsen, Quirine F Manson, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

Michael Gadermayr and Maximilian Tschuchnig. Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, 112:102337, 2024.

Sungrae Hong, Sol Lee, Jisu Shin, Jiwon Jeong, and Mun Yong Yi. Diagnose like a real pathologist: An uncertainty-focused approach for trustworthy multi-resolution multiple instance learning. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6132–6141. IEEE, 2026.

Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. 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.

Peixiang Huang, Yanyan Huang, Weiqin Zhao, Junjun He, and Lequan Yu. Hyperpath: Knowledge-guided hyperbolic semantic hierarchy modeling for wsi analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 262–272. Springer, 2025.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. Pmlr, 2018.

Md Amirul Islam, Sen Jia, and Neil DB Bruce. How much position information do convolutional neural networks encode? *arXiv preprint arXiv:2001.08248*, 2020.

Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3344–3354. IEEE, 2023.

David Kim, Liron Pantanowitz, Peter Schüttler, Dig Vijay Kumar Yarlagadda, Orly Ardon, Victor E Reuter, Meera Hameed, David S Klimstra, and Matthew G Hanna. (re) defining the high-power field for digital pathology. *Journal of pathology informatics*, 11(1):33, 2020.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Maxime W Lafarge, Erik J Bekkers, Josien PW Pluim, Remco Duits, and Mitko Veta. Roto-translation equivariant convolutional networks: Application to histopathology image analysis. *Medical Image Analysis*, 68:101849, 2021.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14313–14323. IEEE, 2021.

Pei Liu, Luping Ji, Jiaxiang Gou, Bo Fu, and Mao Ye. Interpretable vision-language survival analysis with ordinal inductive bias for computational pathology. In *International Conference on Learning Representations*, volume 2025, pp. 99062–99090, 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

Taro Nitta and Masataka Kawai. Rotransmil: Rotary positional embedding for transformer-based multiple instance learning. In *Proceedings of the 2025 5th International Conference on Artificial Intelligence and Application Technologies*, pp. 47–51, 2025.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Linhao Qu, Siyu Liu, Xiaoyu Liu, Manning Wang, and Zhijian Song. Towards label-efficient automatic diagnosis and analysis: a comprehensive survey of advanced deep learning-based weakly-supervised, semi-supervised and self-supervised techniques in histopathological image analysis. *Physics in Medicine & Biology*, 67(20):20TR01, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Jan Ramon, Luc De Raedt, and Stefan Kramer. Multi instance neural networks. In *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*, pp. 53–60, 2000.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pp. 4460–4469. PMLR, 2018.

Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018.

Daniel Shao, Richard J Chen, Andrew H Song, Joel Runevic, Ming Y Lu, Tong Ding, and Faisal Mahmood. Do multiple instance learning models transfer? *arXiv preprint arXiv:2506.09022*, 2025.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, 2018.

Vighnesh Shiv and Chris Quirk. Novel positional encodings to enable tree-based transformers. *Advances in neural information processing systems*, 32, 2019.

Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew FK Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *European Conference on Computer Vision*, pp. 699–715. Springer, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yogesh Verma, Amauri H Souza, and Vikas Garg. Positional encoding meets persistent homology on graphs. *arXiv preprint arXiv:2506.05814*, 2025.

Conghao Xiong, Hao Chen, Joseph JY Sung, and Irwin King. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. *arXiv preprint arXiv:2301.08125*, 2023.

Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in Oncology*, pp. 4133, 2021.

Shu Yang, Yihui Wang, and Hao Chen. 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.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18780–18790. IEEE, 2022.

Yuchen Zhang, Zeyu Gao, Kai He, Chen Li, and Rui Mao. From patches to wsis: A systematic review of deep multiple instance learning in computational pathology. *Information Fusion*, 119: 103027, 2025.

A Detailed Description on the Proposed Method

A.1 Proofs on H2PE

Proposition 1 (Rotation Invariance)**.** *When $m_i = m_j$, d_{ij} is invariant under arbitrary rotations and translations of the coordinate system $c \mapsto \mathcal{R}c + t$, where $\mathcal{R}$ is an orthogonal matrix.*

Proof. Letting $m_i = m_j = m$, Equ. 3 simplifies to $d_{ij} = \operatorname{arccosh}\big(1 + \|c_i - c_j\|^2/(2P^2m^2)\big)$. The right hand side depends solely on the Euclidean norm of $c_i - c_j$ and is independent of direction. Because $\|\mathcal{R}(c_i - c_j)\| = \|c_i - c_j\|$ for any orthogonal matrix $\mathcal{R}$, d_{ij} remains invariant under such rotations and translations. □

Proposition 2 (Path Additivity)**.** *If $c_i = c_j = c_k$ and $m_i \geq m_j \geq m_k$ (or vice versa), then $d_{ik} = d_{ij} + d_{jk}$ holds.*

Proof. When $c_i = c_j$, Equ. 3 simplifies to $d_{ij} = \operatorname{arccosh}\big((r + r^{-1})/2\big)$ for $r = m_i/m_j$, as the path connecting two points with $c_i = c_j$ forms a vertical line in the UHS model, which is a geodesic. If $m_i \geq m_j \geq m_k$, point j lies on this geodesic connecting i and k, making the distances additive. For instance, evaluating at $\times 5$, $\times 10$, and $\times 20$ at the same physical position with adjacent MPP ratios of 2 yields $d(\times 5, \times 10) = d(\times 10, \times 20) = \operatorname{arccosh}(5/4) = \ln 2$ and $d(\times 5, \times 20) = \operatorname{arccosh}(17/8) = \ln 4 = 2 \ln 2$. Consequently, $d(\times 5, \times 20) = d(\times 5, \times 10) + d(\times 10, \times 20)$ holds exactly. □

Proposition 3 (Domain Safety)**.** *The argument of* $\operatorname{arccosh}$ *in Equ. 3 is always greater than or equal to 1 for any arbitrary input.*

Proof. The numerator is a sum of squares and is thus always non negative (≥ 0), while the denominator is strictly positive (> 0) since $m_i, m_j > 0$. Therefore, $1 + (\text{non negative})/(\text{positive}) \geq 1$ holds universally, ensuring that no input falls outside the domain. This implies that the numerical clipping typically required in hyperbolic neural network implementations is completely unnecessary for this formulation. □

Table 2: Time complexity of PEs applied to MRMIL. N is the number of patches, H is the MLP hidden dimension, k is the neighborhood size.

PE	Time complexity	Note
PPEG	$O(N)$	Grid convolution with local kernels
RoPE	$O(N)$	Per token rotation, attention cost separate
TNAM	$O(Nk)$	Local neighbor aggregation
Sin. PE	$O(N)$	Fixed additive function
HyPE	$O(N^2)$	Closed form pairwise bias, no MLP
TreePE	$O(N \log N)$	Root to node path composition
GPSE	$O(N)$	Fixed depth MPNN over sparse graph
H2PE	$O(N^2H)$ dense, $O(NH)$ sparse	Pairwise MLP bias, block fused

A.2 Computational Complexity

d_{ij}, Δ_{ij}, and g_i are deterministic coordinate functions independent of network parameters and are cached once per slide. γ_{ij} depends on the learnable γ_0, β and is recomputed at every training step, but as an elementwise operation on two scalars it is never the bottleneck for either quantity. The actual bottleneck is the evaluation of MLP_r. Unlike $QK^\top$, which reduces directly to an (N, N) score without materializing an (N, N, d_k) intermediate, MLP_r is applied independently to each pair and, naively implemented, requires materializing $(N, N, 2)$ inputs and (N, N, H) hidden activations, which for N in the tens of thousands dominates the memory footprint of dense self-attention itself. We resolve this by fusing the MLP_r evaluation into the same query key block traversal used to compute $QK^\top$, following the same principle FlashAttention applies to $QK^\top$ itself. For block size B, $A^{(p,q)} = Q_p K_q^\top / \sqrt{d_k} \in \mathbb{R}^{B \times B}$ is computed together with $b^{(p,q)}$ at the same (B, B, H) scale, immediately reduced to $A^{(p,q)} - b^{(p,q)}$, and discarded, bounding peak memory at $O(B^2H)$ rather than $O(N^2H)$ while keeping total FLOPs at the same $O(N^2H)$ order dense self-attention already incurs. Message passing GNNs operate on a sparse edge set by construction, reducing this to $O(NH)$. Tab. 2 summarizes the time complexity of H2PE alongside all comparison PEs. Most baselines are cheaper in the worst case. PPEG, RoPE, Sin. PE, and GPSE scale linearly in N since each applies a fixed local operation per token or per edge of a sparse auxiliary graph. TNAM scales with a fixed neighborhood size k. TreePE scales as $N \log N$ through its root to node path composition. HyPE shares H2PE's $O(N^2)$ order, as it likewise forms a bias over all attention pairs, but

Algorithm 1 H2PE Forward Computation

Require: Patch features $\{v_i\}_{i=1}^N$, coordinates $\{(x_i, y_i)\}_{i=1}^N$, MPP values $\{m_i\}_{i=1}^N$, patch pixel size P, pair set E required by the downstream aggregator, learnable $\gamma_0, \beta, \mathrm{MLP}_g, \mathrm{MLP}_r$
Ensure: Updated patch features $\{v_i'\}_{i=1}^N$, pairwise bias $\{b_{ij}\}_{(i,j)\in E}$
Step 1: Per patch position extraction
for $i = 1$ to N **do**
  Compute $z_i \leftarrow m_i \times P$
  Compute $c_i \leftarrow z_i \times (x_i + \frac{1}{2}, y_i + \frac{1}{2})$
  Compute $g_i \leftarrow \ln m_i$
  Compute $e_i \leftarrow \mathrm{MLP}_g(g_i)$
  Update $v_i' \leftarrow v_i + e_i$
end for
Step 2: Pairwise relation and bias computation
Compute $m_{\text{ref}} \leftarrow \max_i m_i$
for each pair $(i, j) \in E$ **do**
  Compute d_{ij} using Equ. 3
  Compute $\Delta_{ij} \leftarrow g_i - g_j$
  Compute $\gamma_{ij} \leftarrow \gamma_0 \cdot (m_{\text{ref}} / \min(m_i, m_j))^\beta$
  Compute $b_{ij} \leftarrow \mathrm{MLP}_r\big(\mathrm{Concat}(d_{ij}, \gamma_{ij}\Delta_{ij})\big)$
end for
return $\{v_i'\}_{i=1}^N$, $\{b_{ij}\}_{(i,j)\in E}$

carries no additional H factor in the absence of an MLP. H2PE is therefore not the asymptotically cheapest method among those compared. In practice, however, this distinction has little bearing on wall clock cost. As shown in Fig. 5(a), H2PE completes position construction in roughly 0.01 seconds per WSI, two orders of magnitude below the 1 second budget within which every method in Tab. 2 already operates. At the patch counts realized in practice, this asymptotic gap does not translate into a meaningful difference in real world runtime.

A.3 Discussion on Alternative Hyperbolic Models

Although the Poincaré ball model also represents hyperbolic space, its isometry group consists of Möbius transformations that fix the origin. While well suited for embeddings placed freely through gradient-based optimization (i.e., taxonomy word embeddings), this property introduces arbitrary choices regarding origin placement when explicitly computing closed-form coordinates from (x_i, y_i, m_i) as in our framework. Conversely, the UHS model permits the direct use of a global coordinate system without prioritizing any specific boundary point, aligning naturally with the underlying coordinate structure of our problem.

A.4 MLP Recipe

Both MLP_g and MLP_r share an identical compact structure consisting of two linear layers with an intervening non linear activation function, differing only in their input and output dimensions:

$$\begin{aligned} \mathrm{MLP}_g &: \mathbb{R}^1 \xrightarrow{\text{Linear}} \mathbb{R}^H \xrightarrow{\text{GELU}} \mathbb{R}^H \xrightarrow{\text{Linear}} \mathbb{R}^d, \\ \mathrm{MLP}_r &: \mathbb{R}^2 \xrightarrow{\text{Linear}} \mathbb{R}^H \xrightarrow{\text{GELU}} \mathbb{R}^H \xrightarrow{\text{Linear}} \mathbb{R}^1 \end{aligned} \tag{7}$$

where H denotes the hidden dimension fixed to 128. The input fed into MLP_g is the scalar g_i, whereas MLP_r receives the two dimensional vector $\mathrm{Concat}(d_{ij}, \gamma_{ij}\Delta_{ij})$ specified in Equ. 5. For both MLPs, the weights and biases of the final linear layer are initialized to zero. This zero initialization ensures $e_i \approx 0$ and $b_{ij} \approx 0$ at the onset of training, allowing the model to behave identically to the baseline aggregator without H2PE while gradually increasing the PE contribution as optimization proceeds.

B Detailed Implementation Details

B.1 Detailed Fusion Specifications across Families

The three MRMIL families expose node level and pairwise interaction at different stages, so the fusion point of each positional encoding differs by family. Every encoding is attached through at most two hooks: a node additive term that is added to the patch features entering the aggregator, and a pairwise term that is consumed through the family specific path described below. PPEG, TNAM, Sinusoidal PE, and RoPE provide only the node additive term. TreePE, GPSE, and HyPE provide both. TreePE adds an embedding accumulated along the quadtree path from the root to each patch and exposes the tree distance between two paths as its pairwise term, GPSE adds random walk structural features on the 4-connected grid graph within each level and exposes a linearly mapped hop count as its pairwise term, and HyPE adds a projection of the hyperbolic norm of each grid coordinate and exposes its Poincaré distance bias as its pairwise term. RoPE is adapted as a node additive encoding for all families, where a linear projection of each patch feature is rotated by angles derived from its 2D grid coordinates and added back, since none of the aggregators expose query and key vectors. H2PE provides e_i as the node additive term and b_{ij} as the pairwise term.

Interaction set. All pairwise terms are consumed on a common sparse interaction set $\mathcal{E}$ shared by the three families. $\mathcal{E}$ contains the 4-neighbor lattice edges within each level, the parent child edges between adjacent MPP levels, and the edges between the coarsest and the finest level that connect each finest patch to the coarsest patch containing it. Every edge is stored in both directions. Optionally, $\mathcal{E}$ is closed under h-hop reachability so that (i, j) is included whenever the two patches are connected by a path of length at most h in the base set; we use $h = 2$ for H2PE on ZoomMIL and H^{2}MIL and $h = 1$ otherwise. For H2PE, b_{ij} is evaluated only on $\mathcal{E}$, and for a directed edge $(j \to i)$ the antisymmetric term is taken along the message direction, i.e., $\Delta_{ji} = \ln(m_j/m_i)$. The pairwise terms of TreePE, GPSE, and HyPE are computed densely and gathered on $\mathcal{E}$.

(i) ZoomMIL. ZoomMIL performs level wise gated attention pooling rather than dense self attention, so no pairwise attention matrix is exposed. The node additive term is added to the projected patch features, after the level embedding and before pooling. The pairwise term enters through the relational residual module of Equ. 6 evaluated on $\mathcal{E}$: for each patch i, the biases of its incoming edges are normalized by a softmax over j, the projected features Wv'_j are averaged accordingly, and the resulting h_i is added to the patch features before level pooling. A single round of message passing is applied. TreePE, GPSE, and HyPE follow the same path with their own pairwise terms.

(ii) H^{2}MIL. The node additive term is added to the input node features, after the level embedding and before graph convolution, and both convolution layers operate on the same $\mathcal{E}$. The pairwise term modulates message passing as an edge weight $w_{ji} = \exp(-[b_{ji}]_+)$ on each directed edge $(j \to i)$, where $[\cdot]_+ = \max(\cdot, 0)$, so that a non positive bias leaves the message unchanged and a larger bias attenuates it. Weighted messages are summed over incoming edges and divided by the in degree. This is the edge weight form of $A'_{ij} = A_{ij} - b_{ij}$ in Sec. 3.2.4. TreePE, GPSE, and HyPE follow the same path.

(iii) DSMIL. DSMIL consumes spatially aligned pillars: each instance is indexed by a patch at the finest MPP and concatenates the features of that patch with those of its ancestors at every coarser level, which are mixed by a learnable softmax weighted sum of per level projections. All instances therefore share the finest MPP and lie on the finest grid, and the positional context of every encoding is this single grid. The node additive term is added at the instance encoding stage, after the level mixing, and is seen by both the instance classifier and the bag classifier. DSMIL computes attention between each instance and the max pooled critical instance rather than between all pairs, so no pairwise score is exposed. For the bag classifier, each instance i receives a scalar bias obtained by averaging b_{ij} over the edges $(i, j) \in \mathcal{E}$ incident to it, which is subtracted from its attention logit before scaling and normalization. TreePE, GPSE, and HyPE follow the same path, with the average taken over all instances since their pairwise terms are dense. The relational residual of Equ. 6 is not used for this family.

B.2 Positional Encoding visualization

Instead of directly projecting vector PEs into color space, we designate a specific anchor patch as a reference point and derive a scalar quantity for each patch reflecting its similarity, distance, or

relational interaction, subsequently colorizing these values via the jet colormap. Within this representation, the x, y, z coordinates of the 3D point cloud strictly serve to illustrate the spatial layout across the MPP shelf, leaving the color channel to exclusively encode the positional information.