MVH-MIL: Computation-Aware WSI Analysis with Hybrid Hyperbolic-Euclidean Embeddings and Dynamic Sparsity

Anonymous CVPR submission

Paper ID 11826

Abstract

Pathological whole slide images (WSIs) are fundamental in pathology analysis, but their massive image scale and subtle morphological features demand highly trained and experienced pathologists. Deep learning techniques have thus been applied to assist pathologists; however, three key challenges remain in WSI analysis: (i) suboptimal geometric representation, as conventional embeddings inadequately model the intrinsic hierarchical tissue ontologies; (ii) a geometric trade-off between capturing global hierarchy context and preserving local discriminative detail, which hampers multi-scale reasoning; and (iii) inefficient, homogeneous computation pipelines that fail to accommodate pathological heterogeneity. Existing methods offer limited improvements, as they rely on a single geometric paradigm or homogeneous computations. We propose MVH-MIL, a novel geometric-aware hierarchical framework for WSI analysis that: (1) introduces hyperbolic embeddings to naturally model the global, tree-like organization of pathological structures; (2) employs a Hybrid Geometric Embedding strategy that synergistically combines hyperbolic representations with Euclidean embeddings; and (3) integrates a Mixture of WSI Experts (MoWE) module that adaptively routes tissue-specific tokens to specialized sub-networks for efficient, expert-driven handling of pathological heterogeneity. Across 7 WSI datasets spanning 7 cancer types, MVH-MIL consistently improves performance over strong state-of-the-art backbones. Extensive experiments demonstrate that MVH-MIL provides a route for a more robust and scalable approach to efficient digital pathology modeling. The project will be open-sourced.

1. Introduction

Accurate cancer diagnosis and subsequent treatment planning hinge on pathological interpretation, the gold standard for identifying disease from microscopic images [11]. With the rise of digital pathology, tissue biopsies are now commonly scanned into gigapixel-scale Whole Slide Images (WSIs) that capture a wealth of multi-scale spatial information. Despite their richness in information, the sheer size of these images (e.g., $60,000 \times 80,000$ pixels) presents a major bottleneck. Manual WSI review is an arduous task demanding highly specialized expertise, which introduces the risk of inconsistent and variable diagnostic outcomes [20]. Furthermore, their massive spatial scale, high morphological heterogeneity, and inherent hierarchical organization (cells to entire slides) pose major barriers to automated analysis [9, 12]. To address these challenges, Multiple Instance Learning (MIL) is adopted as a dominant framework for WSI analysis [24, 33]. It tackles gigapixel scale by partitioning WSIs into small tiles (e.g., 224×224 pixels) and using embeddings to compress the features. In MIL, a foundation model (e.g., CONCH [22]) first converts tile latent representations into embeddings, and a slide-level model (e.g., ABMIL [14]) then aggregates these embeddings for slide-level predictions. Though practical, current MIL methods struggle to balance three critical requirements:

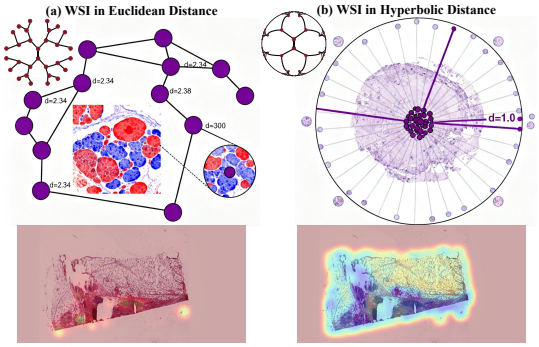

Figure 1. WSI geometric representations. (a) Euclidean embeddings fail to model hierarchical relationships. (b) Hyperbolic geometry naturally encodes the tree-like structure of pathology, from tissue architecture to cellular morphology.

1) ***Hierarchical modeling capability*** [23]. The intrinsic hierarchical nature of WSIs poses challenges for representation in Euclidean space, which is commonly used by traditional slide-level methods [14, 30]. The Euclidean metric alone is suboptimal for capturing complex spatial

CVPR
#11826

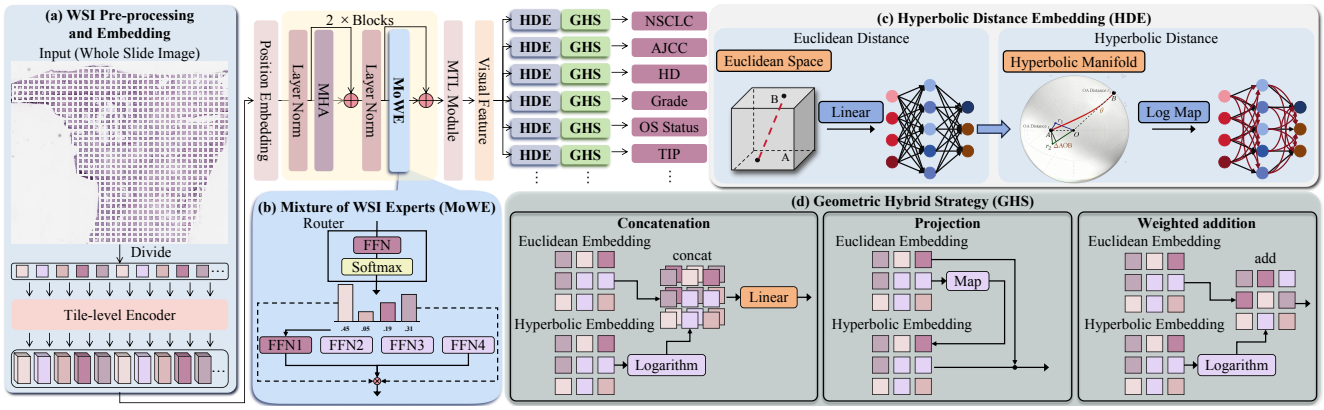

Figure 2. MVH-MIL overview. (a) A WSI is first partitioned into tiles, then extracted into embeddings with a pathological foundation model for the later slide-level modeling. (b) MoWE employs WSI-specialized experts to extract domain-specific features. (c) HDE converts Euclidean representations to hyperbolic distance to capture both hierarchical relationships. (d) GHS combines hyperbolic and Euclidean embeddings to preserve local details and create an integrated, multi-scale representation.

relationships [13] and restricts the capacity of existing approaches to model structural hierarchies, ranging from individual cells to glands and large tissue architecture (Fig. 1).

2) ***Efficient homogeneous computation*** [34]. Features in pathology images are highly heterogeneous. Diagnostic regions of interest are typically small and dispersed within extensive areas of benign or irrelevant tissues. Existing methods [27, 28] apply uniform computation across entire slides, failing to adapt to this variability and wasting substantial resources on trivial regions.

3) ***Multi-scale reasoning*** [25]. WSI analysis requires the simultaneous understanding of high-level tissue architecture and fine-grained cellular morphology. The geometric trade-off limits the ability of current methods [21, 34] to effectively capture global hierarchical context while preserving locally discriminative details. It also impedes the integration of information across spatial scales.

To address these requirements, we propose Mixture Visual Hierarchy MIL (MVH-MIL) (Fig. 2), a novel MIL framework for WSI analysis with three key contributions:

1) We introduce **hyperbolic embeddings modeled in the Poincaré ball** [17, 31] leveraging the unique ability of hyperbolic geometry to represent hierarchical tissue structures. 2) We integrate a **dynamic sparse activation mechanism** into a Vision Transformer (ViT) to enable computation-aware feature extraction. It adaptively focuses computational resources on diagnostically relevant regions, improving efficiency while preserving feature discriminability. 3) We employ a **hybrid embedding strategy** that integrates hyperbolic embeddings with Euclidean embeddings, utilizing Euclidean space to preserve local morphological details and achieve a unified representation of multi-scale contextual information.

Experiments on 7 downstream tasks over 10 state-of-the-art (SOTA) methods across 7 WSI datasets, and prove MVH-MIL's resilience and flexibility. In summary, by integrating pathology-aware inductive biases and employing a targeted hybrid geometric embedding strategy, our framework effectively meets the core requirements in WSI analysis with superior overall performance.

2. Related Works

2.1. Multiple Instance Learning

Deep learning for WSI analysis has evolved toward a two-stage MIL paradigm, separating tile-level feature extraction from slide-level aggregation. Early approaches, such as SlideAve (averaging) and SlideMax (max-pooling), adopted simplistic aggregation schemes that either diluted critical signals or amplified noise. To address this, attention-based methods like ABMIL [14] and CLAM [21] became central, using adaptive weighting to prioritize informative tiles. Further refinements like DTFD-MIL [34] and DSMIL [19] used feature distillation and contrastive learning, respectively. However, these attention-driven models failed to solve the spatial context problem, struggling to capture inter-tile relationships. To solve this spatial limitation, TransMIL [27] used transformers to encode tile relationships. This introduced new issues: poor modeling of hierarchical features and a computational bottleneck from quadratic attention. Research then focused on this bottleneck, splitting into two branches. One branch applies large backbones like LongNet in GigaPath [30], remaining memory-intensive and monolithic. The other branch, State Space Models (SSMs), captures long-range dependencies with linear or sub-quadratic complexity. Representatives includes MambaMIL [32] and PathRWKV [2]. While both branches improve efficiency, they converge on the same unresolved problem: their monolithic designs are ill-equipped to model the diverse and heterogeneous feature distributions found in WSIs.

2.2. Mixture of Experts in Transformers

The Mixture of Experts (MoE) paradigm has re-emerged as a leading approach for scaling Transformer models beyond the limits of dense computation. It replaces the standard dense Feed-Forward Network (FFN) layers with multiple parallel "expert" FFNs. For each input token, a sparse subset of these experts (e.g., the Top-1 or Top-2) are dynamically selected by a lightweight gating network called router. This conditional computation mechanism decouples the model's total parameter count from its computational cost. It allows for a massive increase in model capacity while keeping the computational cost for inference constant. The resurgence was pioneered by GShard-M4 [18], which first successfully applied MoE to scale Transformers for machine translation. Building on this, the Switch Transformer [8] significantly simplified the architecture by using a Top-1 routing strategy, and introduced an auxiliary load balancing loss to ensure stable training and prevent expert collapse. More recently, models like Mixtral 8x7B [15] have demonstrated the power of Top-K routing. They achieve superior performance by combining the outputs of the two best-scoring experts. MoE has also proven effective beyond language, as Vision MoE [26] successfully adapted it to the Vision Transformer by routing individual image patches, confirming MoE's status as a general-purpose scaling solution.

2.3. Hyperbolic Space

Previous MIL methods generally rely on Euclidean geometry. This is considered "geometrically suboptimal" for WSI representation [13], as it fails to capture the intrinsic hierarchical organization of WSIs (Fig. 1a). To better modeling these structures, hyperbolic geometry has been proposed as an effective alternative [13, 17, 31]. Specifically, hyperbolic spaces (e.g., the Poincaré ball [31]) exhibit exponential volume growth, which is well-suited for embedding tree-like, hierarchical data. This matches the features in WSIs (Fig. 1b). Recent works such as HyperPath [13] and HVHM [17] have leveraged hyperbolic neural networks to model these tissue hierarchies, enabling an improved representation of structural organization. However, these methods often face a fundamental geometric trade-off [7, 25]: while hyperbolic space excels at capturing global hierarchy, it distorts fine-grained local and discriminative features that are better preserved in Euclidean space. This trade-off creates the need to jointly model global hierarchy and local details.

3. Methods

3.1. WSI Pre-processing and Embedding

A WSI I is first loaded at a specific microns-per-pixel (mpp) resolution and subsequently tessellated into a non-overlapping grid of tiles $\{T_{i,j}\}$, each of size T_s (Fig. 2a). To ensure the quality of these tiles, a two-step screening process is applied. First, tiles are discarded if their tissue content is below a predefined threshold (e.g., $< 70\%$ area coverage). Second, tiles are eliminated if their pixel variance is beneath a quantitative cutoff (e.g., $Var(I) < 0.05$, with I as the pixel intensity normalized to the [0,1] range). This ensures only informative tiles proceed to the next stage.

Each valid tile is then converted into a compact, semantic feature vector via a pathological foundation model (e.g., GigaPath [30]). This feature extraction step serves to improve both the training efficiency and the overall performance of the downstream slide-level backbone (e.g., TransMIL [27]).

3.2. Mixture of WSI Experts

The slide-level backbone is based on ViT [6], with the major modification on its FFN component. A basic FFN processes the same transformation for each tile embedding h_i:

$$\mathrm{FFN}(h_i) = \mathrm{GELU}(h_i W_1 + b_1)W_2 + b_2 \quad (1)$$

This is incompatible with the heterogeneous nature of WSI, ignoring its diverse semantic properties. To address this, we adapt the concept of Mixture-of-Experts (MoE) [26] to computational pathology by proposing the Mixture of WSI Experts (MoWE) (Fig. 2b). This layer replaces the dense FFN with K specialized expert networks $\{E_1, E_2, ..., E_K\}$, each learning distinct pathological patterns. However, computing all K experts for every token h_i (a "dense" MoE) is computationally prohibitive. Instead, MoWE implements a sparse, Top-K routing strategy. A single, learnable routing network R (implemented as a single linear layer) first computes K gating logits $l_i \in \mathbb{R}^K$ for each token h_i:

$$l_i = \mathrm{R}(h_i) = h_i W_g \quad (2)$$

where W_g is the router's weight matrix. Next, rather than directly implement softmax as MoE, we use a Top-K operation to select the indices I of the Top-K experts:

$$I, l_{i,k} = \text{Top-K}(l_i, k) \quad k \ll K \quad (3)$$

where $I = \{j_1, \ldots, j_k\}$ is the set of chosen expert indices and $l_{i,k}$ are their corresponding logits. We calculate the gating weights w_i by applying softmax only to these k logits:

$$w_i = \mathrm{Softmax}(l_{i,k}) \quad (4)$$

The final output $\mathrm{MoWE}(h_i)$ is the sparsely weighted sum of only the k activated experts:

$$\mathrm{MoWE}(h_i) = \sum_{j \in I} w_{i,j} \cdot E_j(h_i) \quad (5)$$

This sparse approach significantly reduces computational cost, as only k experts are computed. It also maintains high model capacity with a large K.

Additionally, we introduce the auxiliary loss $\mathcal{L}_{\text{aux}}$ during training to prevent expert collapse, where the router overwhelmingly favors a few experts. It encourages the router to distribute tokens evenly across all K experts. This is based on the fraction of tokens f_i routed to expert i, and the mean routing probability P_i for that expert across the batch:

$$\mathcal{L}aux = \alpha \cdot \sum_{i=1}^{K} f_i \cdot P_i \tag{6}$$

where α is a hyperparameter. $\mathcal{L}_{\text{aux}}$ is added to the main model loss and optimized jointly.

3.3. Hyperbolic Distance Embedding

We adopt hyperbolic distance embedding (HDE) (Fig. 2c) on the output from the slide-level backbone, additionally modeling the features from a broader scale. Existing MIL methods [14, 19, 21, 30] often operate in Euclidean space (Fig. 1a), measuring similarity via the L_2 norm:

$$D_{\text{euc}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{7}$$

This geometry excels at modeling local discriminative patterns but misaligns with the intrinsic, multi-scale hierarchy of WSIs. To address this, we adopt hyperbolic space via the Poincaré ball model $\mathbb{M}_c^n$, a principled alternative for embedding hierarchies. Its distance metric D_{hyp} is derived from Möbius addition ($\oplus_c$):

$$D_{\text{hyp}}(\mathbf{x}, \mathbf{y}) = \frac{2}{\sqrt{c}} \text{arctanh}\left(\sqrt{c} \left\| -\mathbf{x} \oplus_c \mathbf{y} \right\|\right) \tag{8}$$

$$\mathbf{x} \oplus_c \mathbf{y} = \frac{(1 + 2c\langle \mathbf{x}, \mathbf{y} \rangle + c\|\mathbf{y}\|^2)\,\mathbf{x} + (1 - c\|\mathbf{x}\|^2)\,\mathbf{y}}{1 + 2c\langle \mathbf{x}, \mathbf{y} \rangle + c^2\|\mathbf{x}\|^2\|\mathbf{y}\|^2} \tag{9}$$

This projection is a form of manifold learning, embedding hierarchical tissue structures from Euclidean space onto a hyperbolic manifold. It scales exponentially ($O(e^{cr})$), unlike polynomial Euclidean growth ($O(r^n)$), preserving their intrinsic geometry.

3.4. Geometric Hybrid Strategy

The global advantage of HDE introduces a local trade-off. Hyperbolic geometry's reliance on a conformal factor, $\lambda_c = \frac{2}{1 - c\|x\|^2}$, can destabilize fine-grained local features. The critical dichotomy of global hierarchical capacity versus local feature preservation motivates our geometric fusion. This is based on the premise that hyperbolic geometry D_{hyp} generalizes Euclidean geometry D_{euc} when $c \to 0$:

$$\lim_{c \to 0} D_{\text{hyp}}(\mathbf{x}, \mathbf{y}) = 2\|\mathbf{x} - \mathbf{y}\| \tag{10}$$

This demonstrates the mathematical continuity between the two geometries. We therefore propose three distinctive fusion strategies:

Concatenation: With identical (z_E, z_H) embeddings, fusion can be performed using concatenation $\oplus$:

$$z_{\text{concat}} = [z_E \oplus \log_0^c(z_H)], \tag{11}$$

this method preserves complete geometric information while expanding dimensionality to $2m$, followed by a linear layer to compress the channel dimension back to m.

Projection: For $h_i, h_j \in \mathbb{R}^d$, we compute $d_E(h_i, h_j) = \|h_i - h_j\|_2$ and their hyperbolic embeddings $z_{H_i} = \exp_0^c(W_H h_i)$, $z_{H_j} = \exp_0^c(W_H h_j)$ in $\mathbb{H}_c^m$. The fused distance d_{f} is:

$$d_{\text{f}}(h_i, h_j) = d_H(z_{H_i}, z_{H_j}) \cdot \left(1 + \gamma \cdot \phi\left(\frac{d_E(h_i, h_j)}{\sigma_E}\right)\right) \tag{12}$$

where $d_H(\cdot, \cdot)$ denotes the Poincaré distance in $\mathbb{H}_c^m$, σ_E is the dataset's Euclidean scale, and γ is a learnable intensity. The projection function $\phi : \mathbb{R} \to [0, 1]$ employs a tempered sigmoid activation, with learnable centering parameter μ and temperature κ controlling the transition sharpness:

$$\phi(x) = \frac{1}{1 + e^{-\kappa(x - \mu)}}, \tag{13}$$

The Poincaré distance exhibits exponential growth with the embedding radius:

$$d_H(z_{H_i}, z_{H_j}) = \text{arcosh}\left(1 + 2\frac{\|z_{H_i} - z_{H_j}\|^2}{(1 - \|z_{H_i}\|^2)(1 - \|z_{H_j}\|^2)}\right) \tag{14}$$

Weighted addition: For a tile-level feature $h \in \mathbb{R}^d$, the hybrid embedding (z_E, z_H) combines $z_E = W_E h \in \mathbb{E}^m$ and $z_H = exp_0^c(W_H h) \in \mathbb{H}_c^m$, where $exp_0^c(\text{v}) = \tanh(\sqrt{c}\frac{\|\text{v}\|}{2})\frac{\text{v}}{\sqrt{c}\|\text{v}\|}$. The fused representation is:

$$z = \alpha \cdot z_E + (1 - \alpha) \cdot \log_0^c(z_H), \tag{15}$$

where α is a learnable gate, balancing the geometries via the inverse log mapping ($\log_0^c$). Element-wise summation preserves dimensionality for cross-geometric refinement.

These strategies represent three distinct fusion philosophies. In our empirical evaluations, weighted addition was ultimately adopted as the geometric hybrid strategy (GHS) (Fig. 2d). It avoids deferring the complex integration burden to a subsequent linear layer, as concatenation does. It also doesn't limit the fusion's scope purely to distance computation, as projection does. Instead, this strategy projects the hyperbolic representation carrying hierarchical context back to its tangent space via the logarithmic map ($\log_0^c$). It also performs a direct, learnable interpolation with the Euclidean representation carrying local details. This process

yields a unified feature representation that most effectively coordinates the strengths of both geometries. It aligns perfectly with the intrinsic multi-scale nature of WSI analysis.

4. Experiment

4.1. Datasets and Downstream Tasks

We validate MVH-MIL on 7 core downstream tasks across 7 WSI datasets, covering diverse diagnostic scenarios in computational pathology (Fig. 3). The **CAMELYON16** [1] dataset with the **Breast Metastasis (BrMet)** task classifies lymph nodes as normal or tumorous. The **TCGA** [4] datasets evaluate multiple diverse diagnostic tasks. Specifically, **TCGA-Lung** performs classification of subtypes of **non-small cell lung cancer (NSCLC)** and predicts the **AJCC-defined tumor staging (T-Stage)**, a metric reflecting the tumor's local progression. For **Histological Diagnosis (HistoDx)** tasks, **TCGA-BRCA** classifies breast cancer subtypes and **TCGA-GBM** diagnoses brain cancer based on morphological patterns. Meanwhile, **TCGA-BRCA** predicts **HER2 (IHC-HER2)** expression status from WSIs to assess suitability for targeted therapy. For **Cancer Grading (Grade)** tasks, **TCGA-BLCA** and **TCGA-CESC** perform on bladder and cervical tissues respectively, determining tumor differentiation levels. And **TCGA-CESC** dataset with the **Lymphovascular Invasion (LymInv)** task identifies the presence of cancer invasion into lymphatic or vascular channels. The **TCGA-LGG** dataset with the **T-Stage** task predicts brain tumors staging.

4.2. Implementation Details

All models were uniformly initialized without pre-training. The training regimen included a 20-epoch linear warm-up, followed by 80 epochs optimized via Adam [16], with the learning rate decayed to $0.01\times$ its initial value using a cosine schedule. We adopted hyperparameter search over three orders of magnitude (1×10^{-4}, 1×10^{-5}, and 1×10^{-6}), and reported the best performance for each model. We set $Batchsize = 4$, with $Samp_{Max} = 2,000$ per WSI. Experiments were conducted on 4 NVIDIA RTX 4090 and 2 NVIDIA H100, using Python 3.10.16, PyTorch 2.4.0, and CUDA 12.1. The pipeline is supported by UnPuzzle [20].

4.3. Comparison with SOTA Methods

We compared MVH-MIL against 2 baselines: **SlideAve** and **SlideMax**, and 8 SOTA methods: **ABMIL** [14], **CLAM** [21], **DSMIL** [19], **TransMIL** [27], **LongNet** [5], **GigaPath** [30], **S4MIL** [10], and **MambaMIL** [32]. The comparison is conducted on 7 challenging pathological tasks across 7 distinct datasets, which cover diverse clinical objectives (Tab. 1). This comprehensive evaluation demonstrates the generalizability and robustness of our proposed method.

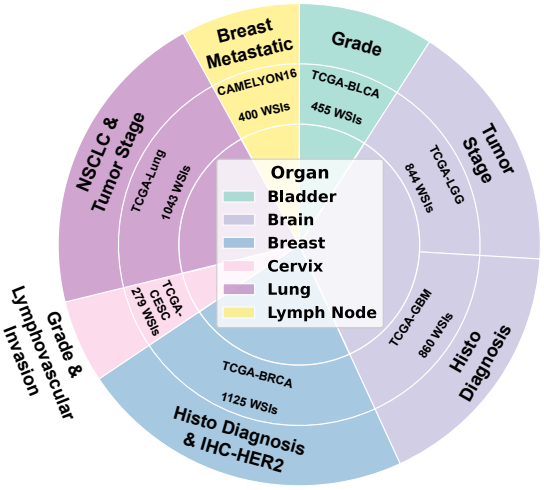

Figure 3. The implemented datasets and downstream tasks.

The results clearly highlight the superiority of MVH-MIL. Our method achieves SOTA performance, ranking first on 7 out of the 7 tasks. Specifically, MVH-MIL achieves the best scores on CAMELYON16 BrMet (96.9%), TCGA-BRCA HistoDx (86.4%), TCGA-BRCA IHC-HER2 (63.4%), TCGA-CESC GRADE (56.1%), and TCGA-GBM HistoDx (99.3%). It also achieves the joint-best performance on TCGA-Lung NSCLC (75.7%, tied with SlideMax) and TCGA-BLCA GRADE (97.7%, tied with LongNet). Furthermore, for the remaining 3 tasks where it does not rank first, MVH-MIL consistently secures the second-best performance, as indicated by the underlined scores for TCGA-Lung T-Stage (50.9%), TCGA-LGG T-Stage (77.9%), and TCGA-CESC LymInv (70.4%). This consistent high performance across all evaluated tasks underscores the model's robustness. Notably, MVH-MIL significantly outperformed all other SOTA methods ($p < 0.05$) in the Wilcoxon signed-rank test [29], with the only exception being S4MIL ($p = 0.064$). This result validates the statistical significance of our approach.

4.4. Visualization Analysis

To understand where different methods focus, we visualize Class Activation Maps (CAMs) using Grad-CAM [3]. We compare the CAMs of three MVH-MIL variants (MVH-MIL$_{cat}$, MVH-MIL$_{emb}$, MVH-MIL$_{add}$) with Euclidean-based methods (Fig. 4). The results show a clear distinction in focus. Most Euclidean-based methods, such as ABMIL and GigaPath, produce overly smooth and diffuse heatmaps that create a halo effect over the entire tissue. Others, like DSMIL and S4MIL, fail to pinpoint specific critical regions, showing minimal or no meaningful activation.

In contrast, all the MVH-MILs generate more interpretable CAMs. They successfully identify multi-scale spatial relationships, as their heatmaps are not one uniform, monolithic blob. Instead, they highlight both broad con-

CVPR #11826

Method	TCGA-BRCA HistoDx	TCGA-GBM HistoDx	TCGA-Lung T-Stage	TCGA-LGG T-Stage	TCGA-BLCA Grade	TCGA-CESC Grade	TCGA-Lung NSCLC	CAMELYON16 BrMet	TCGA-BRCA IHC-HER2	TCGA-CESC LymInv	P-value
SlideAve	81.0	97.2	46.3	74.5	96.6	52.6	75.1	69.6	61.2	63.0	0.002
SlideMax	80.1	97.9	50.0	70.5	95.6	47.4	**75.7**	96.1	57.9	59.3	0.008
ABMIL	83.3	97.9	**51.8**	77.9	96.6	54.4	71.6	93.7	62.8	66.7	0.015
CLAM	78.7	94.4	45.4	75.8	95.5	54.4	72.2	73.4	54.1	66.7	0.002
DSMIL	81.9	97.9	50.0	75.2	96.6	50.9	73.4	86.1	59.0	63.0	0.002
TransMIL	81.9	98.6	50.9	74.5	96.6	50.9	73.4	93.7	59.0	70.4	0.011
LongNet	80.5	94.5	45.4	75.8	**97.7**	50.5	74.6	96.2	55.2	66.7	0.008
GigaPath	84.2	98.6	50.0	75.8	97.3	50.9	75.1	95.3	62.8	66.7	0.002
S4MIL	81.4	98.6	**51.8**	73.8	96.6	52.6	74.6	93.8	58.5	**74.1**	0.064
MambaMIL	83.3	97.9	50.0	**79.2**	95.5	50.9	71.0	95.3	59.6	63.0	0.006
MVH-MIL	**86.4**	**99.3**	50.9	77.9	**97.7**	**56.1**	**75.7**	**96.9**	**63.4**	70.4	-

Table 1. The performance comparison with SOTA methods. **Bold**: the best result. Underline: the second best result.

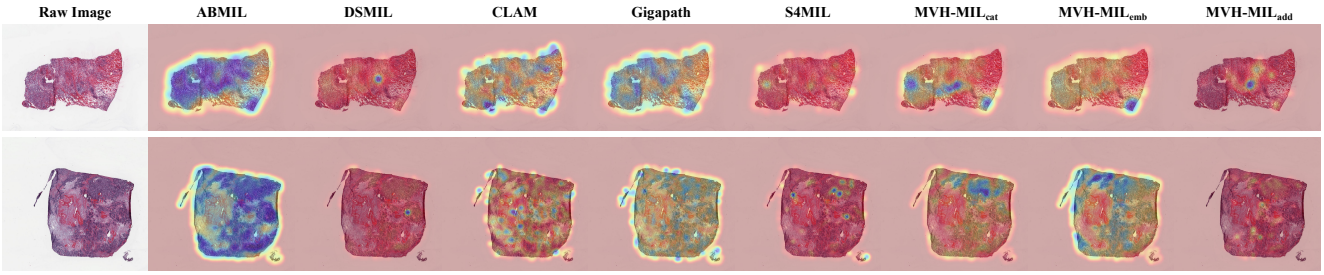

Figure 4. Grad-CAM visualization for different MIL methods. Compared to other Euclidean space-based methods, all three MVH-MIL methods accurately identify the multi-scale spatial relationships among tiles.

textual regions and specific, high-attention focuses. This structured attention demonstrates a more sophisticated understanding of the WSI. Notably, MVH-MIL$_{add}$ provides the most refined visualization, with clean boundaries and a superior capability to seek pathologically significant areas.

5. Discussion

In this section, we evaluate the effectiveness of each proposed module. We first take out a baseline, SlideViT, based on ViT [6]. We modify it to fit as the slide-level backbone. Specifically, the patch embedding module is replaced with a linear layer to directly take the tile embeddings. The original 1D position embedding is also replaced with a 2D Sin-Cos position embedding from GigaPath [30] to embed the coordinates of tiles. We then replace its original FFN module with our proposed MoWE module, or individually add the GHS module to evaluate their individual contributions. These revisions are named SlideMoWE and SlideGHS.

5.1. Mixture of WSI Experts Analysis

5.1.1. Module Ablation

We first evaluate the effectiveness of MoWE (Tab. 2). The results show that MoWE is a crucial and highly effective component for improving model performance. It outperforms the SlideViT in 8 out of 7 downstream tasks, with notable gains in TCGA-Lung T-Stage (49.1% vs. 47.7%) and TCGA-LGG T-Stage (77.2% vs. 75.8%). This demonstrates that the dynamic expert routing mechanism in MoWE is more powerful and adaptive than the static FFN. This effectiveness is further validated by the training and validation curves in Fig. 6. Although the training loss of SlideMoWE is relatively high during comparison, it exhibits the most stable validation curve among all compared methods, showing strong generalization and excellent resistance to overfitting. Furthermore, the radar chart (Fig. 5i) confirms that SlideMoWE achieves a broad and competitive performance profile across diverse tasks. However, the above results also demonstrate the weakness of individually implementing MoWE. SlideMoWE is beaten by SlideViT in 3 of 10 experiments, demonstrating its instability.

5.1.2. Hyperparameter Evaluation

We then compare different hyperparameter settings in MoWE to evaluate their influence on the model performance (Tab. 3). Specifically, we adopt three sets of number of experts N ($N = 4, 8, 16$) in MoWE, each with corresponding Top-Ks (Top-K $< N$). Results on three benchmark datasets and downstream tasks demonstrate that Top-K $= 2$ often yields strong results (with maximum +1.8%, +2.1%, +2.3% on TCGA-Lung, TCGA-BLCA, and CAMELYON16, compared to the lowest score in each task), balancing performance and efficiency. However, its performance is not consistent across different N. Specifically, the Top-K $= 2, N = 4$ setting achieves the best result on CAMELYON16 (93.8%), yet it underperforms compared to $N = 8$ and $N = 16$ settings on TCGA-BLCA (96.6% vs. 97.7%). This may be due to the different WSI processing strategies in different datasets, and different attention requirements for different downstream tasks.

On the other hand, N also shows a complex relationship with performance. A larger N does not consistently lead to

Method	TCGA-BRCA HistoDx	TCGA-GBM HistoDx	TCGA-Lung T-Stage	TCGA-LGG T-Stage	TCGA-BLCA Grade	TCGA-CESC Grade	TCGA-Lung NSCLC	CAMELYON16 BrMet	TCGA-BRCA IHC-HER2	TCGA-CESC LymInv	P-value
SlideViT	80.5	97.2	47.7	75.8	**97.7**	47.4	74.0	93.7	56.3	66.7	0.008
SlideMoWE	81.9	97.2	49.1	77.2	96.6	45.6	74.6	93.8	56.8	63.0	0.002
SlideGHS	83.3	97.9	44.5	75.8	96.6	45.6	74.0	92.2	57.9	59.3	0.002
MVH-MIL	**86.4**	**99.3**	**50.9**	**77.9**	**97.7**	**56.1**	**75.7**	**96.9**	**63.4**	**70.4**	-

Table 2. Ablations on the proposed MoWE and GHS modules. **Bold**: the best result. Underline: the second best result.

Top-K	TCGA-Lung NSCLC			TCGA-BLCA Grade			CAMELYON16 BrMet		
	N=4	N=8	N=16	N=4	N=8	N=16	N=4	N=8	N=16
2	**74.6**	**74.6**	**74.6**	96.6	**97.7**	**97.7**	**93.8**	91.5	91.5
4	-	**74.6**	72.8	-	**97.7**	96.6	-	93.0	91.5
8	-	-	72.8	-	-	95.6	-	-	93.0

Table 3. Ablations on different number of experts and Top-K.

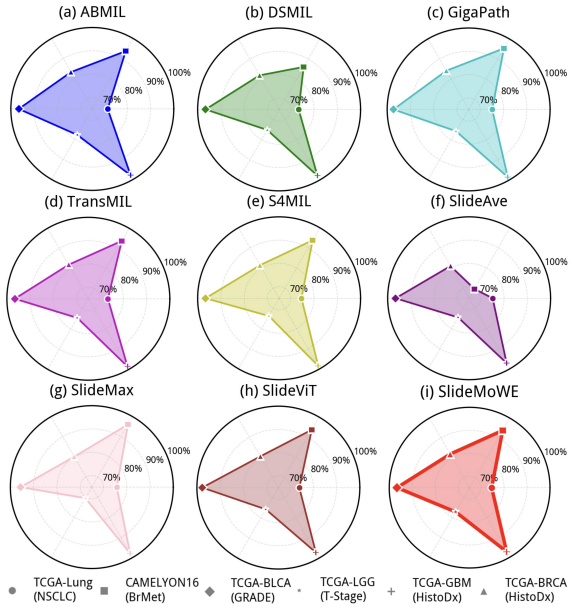

Figure 5. Performance comparison radar charts.

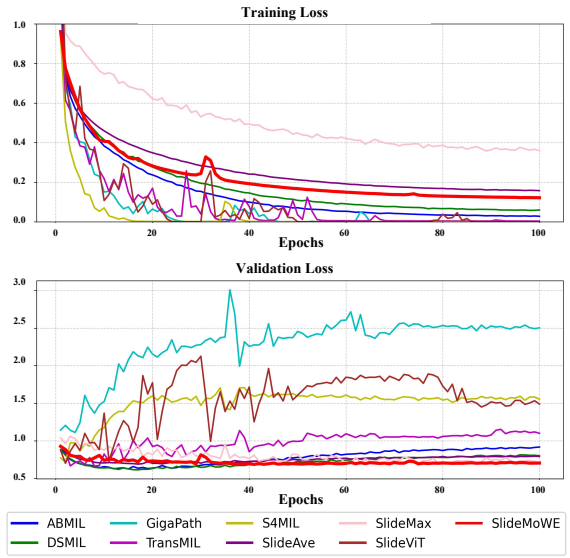

Figure 6. Training and validation loss comparison.

better results. For instance, when Top-K $= 2$, increasing N from 4 to 8 improves the score on TCGA-BLCA (from 96.6% to 97.7%), but significantly degrades it on CAMELYON16 (from 93.8% to 91.5%). For TCGA-Lung, the performance remains identical (74.6%) regardless of N. Furthermore, when Top-K $= 4$, increasing N from 8 to 16 results in a performance drop across all three datasets. This suggests that simply adding more experts may not be an effective strategy and can even lead to inefficient expert specialization, especially when a larger K is also used.

5.2. Hyperbolic Distance Embedding Analysis

Manifold curvature hyp_c is the critical component that determines the geometric structure in the Poincaré ball model. To investigate how hyp_c influences the performance of the MVH-MIL variants, we conduct experiments using 4 different values for this hyperparameter ($hyp_c \in \{0.05, 0.075, 0.1, 0.2\}$). Notably, as $hyp_c \to 0$, the hyperbolic space flattens and collapses to Euclidean space; a larger hyp_c preserves more hyperbolic features in contrast.

Results reveal that the optimal hyp_c value is not universal but is tied to dataset structural properties and variant-specific inductive biases (Tab. 4). Specifically, $hyp_c = \{0.05, 0.075, 0.1, 0.2\}$ achieve 8, 14, 14, and 9 optimal performances on 30 experiments, respectively. Though dataset and variant-specific, moderate hyp_c values ($hyp_c = 0.075, 0.1$) tend to show better generalizability. $hyp_c = 0.1$ achieves the best overall performance, with a slight average performance boost of 0.2% compared with $hyp_c = 0.075$ (75.19% vs. 74.99%). We eventually adopt $hyp_c = 0.1$ based on its superior overall performance.

5.3. Geometric Hybrid Strategy Analysis

5.3.1. Module Ablation

The fusion strategy is critical for deciding the combination structure of the different geometry distances. We evaluate the effectiveness of GHS, compare the performance of three fusion strategies, and further evaluate the generalizability of GHS on other methods. We first evaluate the effectiveness of the GHS module (Tab. 2). The results indicate that implementing GHS individually provides mixed and often unstable results. While it outperforms the SlideViT baseline in 3 out of 10 tasks, with notable gains in TCGA-BRCA

Method	hyp_c	TCGA-BRCA HistoDx	TCGA-GBM HistoDx	TCGA-Lung T-Stage	TCGA-LGG T-Stage	TCGA-BLCA Grade	TCGA-CESC Grade	TCGA-Lung NSCLC	CAMELYON16 BrMet	TCGA-BRCA IHC-HER2	TCGA-CESC LymInv
MVH-MIL$_{cat}$	0.05	**81.9**	97.9	47.2	**75.8**	95.5	50.9	71.0	93.8	**61.7**	71.4
	0.075	**81.9**	**99.3**	46.8	73.2	96.6	47.4	**73.4**	93.8	58.5	70.4
	0.1	81.4	96.5	**53.7**	71.1	95.5	45.6	71.0	**95.3**	57.9	**74.1**
	0.2	81.4	98.6	50.9	72.5	**97.7**	**56.1**	71.0	94.6	60.7	**74.1**
MVH-MIL$_{emb}$	0.05	81.9	97.9	48.2	72.5	95.5	47.4	72.8	89.1	58.5	60.4
	0.075	82.4	**98.6**	**49.1**	73.2	**97.7**	**49.1**	**74.6**	**95.3**	**61.2**	66.7
	0.1	**82.8**	**98.6**	**49.1**	**79.2**	96.6	**49.1**	**74.6**	**95.3**	**61.2**	66.7
	0.2	80.5	**98.6**	48.2	74.5	95.5	**49.1**	**74.6**	**95.3**	56.8	**70.4**
MVH-MIL$_{add}$	0.05	**86.4**	97.9	47.7	73.2	**97.7**	**56.1**	**75.7**	92.2	**63.4**	63.0
	0.075	83.3	**99.3**	50.0	75.2	**97.7**	**56.1**	72.2	96.1	60.1	**70.4**
	0.1	82.4	98.6	**50.9**	**77.9**	96.6	**56.1**	71.6	**96.9**	59.0	**70.4**
	0.2	81.4	97.9	50.0	76.5	96.6	54.4	71.6	**96.9**	61.7	66.7

Table 4. Ablations on different fusion methods with different hyp_c.

(83.3% vs. 80.5%) and TCGA-GBM (97.9% vs. 97.2%), it is outperformed by SlideViT in 5 of the 10 tasks. This demonstrates that hierarchical modeling can be beneficial for specific tasks but is not a universally stable solution on its own, underscoring the necessity of combining GHS with an adaptive component like MoWE to effectively balance hierarchical and Euclidean feature representations.

5.3.2. Strategy Comparison

Different fusion strategies also show distinctive performance (Tab. 4). Specifically, the overall performance of MVH-MIL$_{cat}$ is moderate, securing 5 best scores across the ten experiments, but its unpredictable preference suggests a potential incompatibility with the Poincaré ball model. Besides, MVH-MIL$_{emb}$ only achieves one best score, showing the poorest performance of the three fusion methods. Corresponding to our analysis, MVH-MIL$_{add}$ achieves six best scores across the ten experiments. This result demonstrates its clear superiority, aligning perfectly with the theoretical underpinnings of the Poincaré ball model. Using Möbius addition ($\oplus_c$) is the native, geometrically-aware operation for vector combination within hyperbolic space. This operation is natively compatible with the Poincaré distance metric used for optimization.

5.3.3. Generalization Evaluation

We further evaluate the generalizability of GHS as a plug-and-play module by applying it to seven MIL models (Tab. 5). The results demonstrate that GHS can significantly boost the performance of several backbones, although its impact is task and method dependent. On CAMELYON16, GHS provides dramatic performance gains for specific methods. For example, it improves CLAM from 73.4% to 90.7% (+17.3%) and DSMIL from 86.1% to 94.6% (+8.5%). Similarly, on TCGA-Lung, GHS consistently enhances performance for most methods, notably lifting SlideAve from 75.1% to 78.1% (+3.0%) and ABMIL from 71.6% to 74.6% (+3.0%). On TCGA-BLCA, where performance is already high, the improvements are more modest but still present, as seen with CLAM (95.5% to 97.7%) and ABMIL (96.6% to 97.7%). However, we also observe cases where GHS does not provide a benefit, or even degrades performance, such as with S4MIL on TCGA-Lung (74.6% to 73.4%) and SlideMax on CAMELYON16 (96.1% to 75.2%). This variability suggests that GHS is most impactful for methods and tasks where the original backbone struggles with slide heterogeneity, and its feature-constraining nature may not be universally beneficial for all architectures, requiring a compatible structure design.

Method	TCGA-Lung NSCLC		TCGA-BLCA Grade		CAMELYON16 BrMet	
	o/GHS	w/GHS	o/GHS	w/GHS	o/GHS	w/GHS
SlideAve	75.1	**78.1**	**96.6**	**96.6**	**69.6**	66.7
SlideMax	75.7	**77.5**	**95.6**	94.3	**96.1**	75.2
ABMIL	71.6	**74.6**	96.6	**97.7**	93.7	**94.6**
CLAM	**72.2**	**72.2**	95.5	**97.7**	73.4	**90.7**
DSMIL	73.4	**76.3**	**96.6**	**96.6**	86.1	**94.6**
GigaPath	75.1	**76.3**	97.3	**97.7**	**95.3**	89.9
S4MIL	**74.6**	73.4	**96.6**	**96.6**	**93.8**	92.0

Table 5. Ablations on the mainstream methods with/without GHS. o/GHS is the original method. w/GHS implements GHS.

6. Conclusion

WSI analysis confronts three key challenges: suboptimal tissue hierarchy representation, global-local feature trade-off, and tissue heterogeneity, which our MVH-MIL framework addresses. For hierarchical modeling, we leverage hyperbolic geometry, its exponential distance growth efficiently encodes tree-like pathological structures, reducing Euclidean space information loss. A hybrid geometric embedding module balances scales by fusing Euclidean (local pattern-preserving) and hyperbolic (global hierarchy-capturing) embeddings via weighted addition. To tackle heterogeneity, the MoWE module replaces uniform FFNs with K pathology-specialized experts, using a learnable router to dynamically assign tokens to optimal experts for adaptive computation. MVH-MIL achieves SOTA on 7 downstream tasks across 7 datasets, outperforming 10 SOTA methods. Furthermore, the proposed GHS module and MoWE module are model agnostic and could potentially be adapted to improve other deep learning models, effectively tackling hierarchical and heterogeneous data.

CVPR #11826

References

[1] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, 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. 5

[2] Sicheng Chen, Tianyi Zhang, Dankai Liao, Dandan Li, Low Chang Han, Yanqin Jiang, Yueming Jin, and Shangqing Lyu. Pathrwkv: Enabling whole slide prediction with recurrent-transformer. *arXiv preprint arXiv:2503.03199*, 2025. 2

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017. 5

[4] Antonio Colaprico, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, Isabella Castiglioni, et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, 44(8):e71–e71, 2016. 5

[5] Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023. 5

[6] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 6

[7] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7409–7419, 2022. 3

[8] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 3

[9] Ruiwei Feng, Xuechen Liu, Jintai Chen, Danny Z Chen, Honghao Gao, and Jian Wu. A deep learning approach for colonoscopy pathology wsi analysis: accurate segmentation and classification. *IEEE Journal of Biomedical and Health Informatics*, 25(10):3700–3708, 2020. 1

[10] Leo Fillioux, Joseph Boyd, Maria Vakalopoulou, Paul-Henry Cournède, and Stergios Christodoulidis. Structured state space models for multiple instance learning in digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 594–604. Springer, 2023. 5

[11] Matthew G Hanna, Anil Parwani, and Sahussapont Joseph Sirintrapun. Whole slide imaging: technology and applications. *Advances in Anatomic Pathology*, 27(4):251–259, 2020. 1

[12] Petr Holub, Heimo Müller, Tomáš Bíl, Luca Pireddu, Markus Plass, Fabian Prasser, Irene Schlünder, Kurt Zatloukal, Rudolf Nenutil, and Tomáš Brázdil. Privacy risks of whole-slide image sharing in digital pathology. *Nature Communications*, 14(1):2577, 2023. 1

[13] 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*, pages 262–272. Springer, 2025. 2, 3

[14] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 2, 4, 5

[15] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 3

[16] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[17] Hyeongjun Kwon, Jinhyun Jang, Jin Kim, Kwonyoung Kim, and Kwanghoon Sohn. Improving visual recognition with hyperbolical visual hierarchy mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17364–17374, 2024. 2, 3

[18] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. 3

[19] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 2, 4, 5

[20] Dankai Liao, Sicheng Chen, Nuwa Xi, Qiaochu Xue, Jieyu Li, Lingxuan Hou, Zeyu Liu, Chang Han Low, Yufeng Wu, Yiling Liu, et al. Unpuzzle: A unified framework for pathology image analysis. *arXiv preprint arXiv:2503.03152*, 2025. 1, 5

[21] 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. 2, 4, 5

[22] 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. 1

[23] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 1

[24] Linhao Qu, Manning Wang, Zhijian Song, et al. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *Advances in Neural Information Processing Systems*, 35:15368–15381, 2022. 1

[25] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27263–27272, 2024. 2, 3

[26] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. 3

[27] 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. 2, 3, 5

[28] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–195. Springer, 2021. 2

[29] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945. 5

[30] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, and Yu Gu. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):22, 2024. 1, 2, 3, 4, 5, 6

[31] Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Unsupervised hyperbolic metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12465–12474, 2021. 2, 3

[32] 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*, pages 296–306. Springer, 2024. 2, 5

[33] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical image analysis*, 65: 101789, 2020. 1

[34] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18802–18812, 2022. 2