DiagMIL: Uncertainty-Gated Coarse-to-Fine Multi-Resolution MIL for WSI Diagnosis

Jiwon Jeong[1], Sungrae Hong[1], Donghee Han[1], Jisu Shin[1], Kyungeun Kim[2], and Mun Yong Yi[1]*

[1] Korea Advanced Institute of Science and Technology, Daejeon, South Korea
{zzioni, sr5043, handonghee, jisu3389, munyi}@kaist.ac.kr
[2] Seegene Medical Foundation, Seoul, South Korea
kekim@mf.seegene.com

Abstract. Most advances in whole-slide image (WSI) classification have followed the multiple instance learning (MIL), with multi-resolution (MR) approaches gaining attention under the assumption that higher magnification provides richer evidence. However, high-resolution is not always beneficial: when low magnification already provides sufficient signal, high-magnification details can introduce noise and additional computational cost. This is also inconsistent with clinical practice, where high magnification is typically consulted when the low-magnification overview is inconclusive. Motivated by the coarse-to-fine diagnostic workflow of pathologists, we propose DiagMIL, which reframes MR-WSI analysis for multi-class diagnosis as deciding *when* to consult high-resolution information. DiagMIL performs low-resolution coarse screening with independent class-wise sigmoid evidence first, and our novel differentiable uncertainty-aware gate selectively activates a high-resolution softmax stage only when coarse evidence is uncertain. To support this decision making, DiagMIL leverages structure-aware graph representations with cross-resolution links and structure-level distillation. Extensive experiments show that DiagMIL consistently outperforms strong baselines. Comprehensive analyses validate our perspective on high-resolution integration, showing that selectively invoking fine-grained evidence is more effective and substantially reduces computational cost. Code is available at https://github.com/zzzzzioni/DiagMIL.

Keywords: Digital Pathology · Whole Slide Image · Multiple Instance Learning.

1 Introduction

Gigapixel-scale whole-slide images (WSIs) have become a cornerstone of automated computational pathology diagnosis [8]. However, due to the sheer magnitude of the WSI, it is practically infeasible for pathologists to provide pixel-wise annotations. Therefore, clinicians are often restricted to assigning slide-level labels for training deep learning (DL) models [28]. In such a scenario, Multiple

* Corresponding author.

Instance Learning (MIL) has emerged as a predominant paradigm, which aggregates unlabeled patch-level features using only slide-level supervision [2].

Recent advances in MIL are gradually converging towards multi-resolution (MR) approaches [10]. By leveraging the rich, hierarchical information across multiple scales, MR-MIL frameworks operate through patch pyramids [19], graph structure [12, 14], and multi-scale strategies [7, 13]. Beyond simply utilizing richer information, MR approaches achieve superior performance and interpretability (e.g., attention map) by emulating the diagnostic workflows of pathologists [5].

Despite their promise, existing MR-MILs often integrate multi-scale information in an indiscriminate, always-on manner, reflecting the assumption that high magnification provides richer information [13, 14, 23]. However, high-resolution cues are not always beneficial: when low magnification already provides sufficient signal, additional high-magnification detail can introduce unnecessary noise while substantially increasing processing cost due to the growing patch count [22, 26]. This also diverges from clinical practice, where pathologists often make an early diagnosis when the low-magnification overview provides clear evidence and zoom in only when the coarse assessment is inconclusive to differentiate competing hypotheses [11, 17, 22, 9]. As a result, current MR-MIL pipelines struggle to leverage costly high-magnification evidence in an efficient and diagnostically faithful manner.

Toward overcoming this limitation, we propose **Diag**nosis-inspired dual-stage **MIL** (DiagMIL), an uncertainty-gated framework that performs MR-WSI diagnosis in a selective coarse-to-fine manner. DiagMIL performs coarse screening in low resolution with independent class-wise sigmoid predictions, then activates a high-resolution stage only when the coarse evidence is inconclusive. To drive this routing, we interpret the coarse output as diagnostic evidence and convert it into a Dirichlet belief [20], which provides a principled view of evidence scarcity and class competition. While an uncertainty-aware gate learns when to route a slide to a higher resolution, the fine stage makes a softmax multi-class decision. Furthermore, we model WSIs as structure-aware graphs and enforce cross-resolution consistency through structure-level distillation to support robust gating and prediction. Extensive experiments show that DiagMIL consistently outperforms strong MR-MIL baselines. Comprehensive gating analyses further show that invoking high-resolution cues only when needed both substantially reduces unnecessary fine-scale computation and yields clearer decision boundaries.

2 Method

2.1 Graph and Structure Construction

Each WSI is partitioned into $N^{(r)}$ non-overlapping patches $\{x_i^{(r)}\}_{i=1}^{N^{(r)}}$ at each resolution $r \in \{r_{\text{low}}, r_{\text{high}}\}$, and encoded into patch embeddings $\mathbf{f}^{(r)} = \{f_i^{(r)}\}_{i=1}^{N^{(r)}}$ with coordinates $\mathbf{u}^{(r)} = \{u_i^{(r)}\}_{i=1}^{N^{(r)}}$ by a pre-trained encoder [6]. We then build a patch-level graph $G^{(r)} = (V^{(r)}, E^{(r)})$, where nodes are $\mathbf{f}^{(r)}$ and the edges connect spatial neighbors. We utilize cross-resolution edges between the corresponding

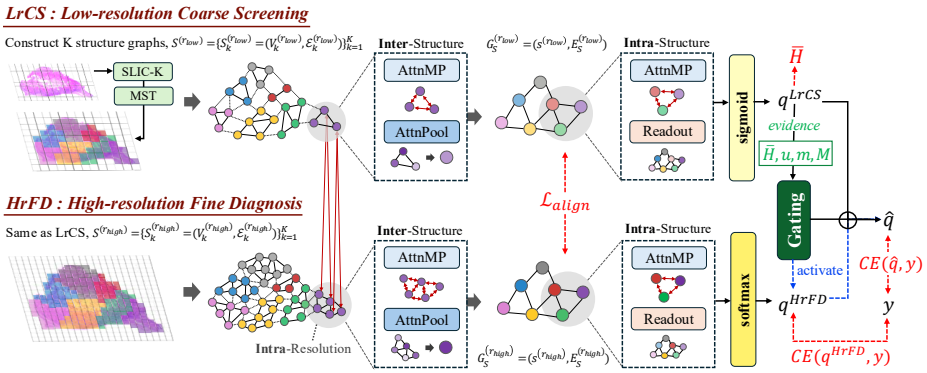

Fig. 1: Overview of the proposed DiagMIL.

low- and high-resolution regions to form $G^{(r_{\text{high}}\leftrightarrow r_{\text{low}})}$. To reflect the structural organization of tissue, we partition each graph $G^{(r)}$ into $K=\lfloor \lambda N^{(r_{\text{low}})} \rfloor$ structures $S^{(r)}=\{S_k^{(r)}\}_{k=1}^{K}$ using the SLIC-K [1] with a joint feature–spatial distance:

$$D(i,k)=\|f_i^{(r)}-\mu_k^{(r)}\|_2^2+\left(\sqrt{\lambda}\|u_i^{(r)}-\nu_k^{(r)}\|_2\right)^2,\quad \lambda\in(0,1), \tag{1}$$

where $\mu_k^{(r)}$ and $\nu_k^{(r)}$ denote the feature and coordinate centroids of structure $S_k^{(r)}$, respectively. Each patch $f_i^{(r)}$ is assigned to the structure $S_k^{(r)}$ minimizing $D(i,k)$, which groups adjacent patches with similar patterns into compact, contiguous tissue regions. To ensure topological connectivity, we refine each structure by extracting a minimum spanning tree (MST) [18] from its edge set E_k:

$$\mathcal{E}_k^{(r)}=\underset{E_k^{\text{MST}}\subseteq E_k}{\arg\min}\sum_{(p,q)\in E_k^{\text{MST}}}\left(\|f_p^{(r)}-f_q^{(r)}\|_2^2+\|u_p^{(r)}-u_q^{(r)}\|_2^2\right). \tag{2}$$

We form K structure graphs $S^{(r)}=\{S_k^{(r)}=(V_k^{(r)},\mathcal{E}_k^{(r)})\}_{k=1}^{K}$ for $r\in\{r_{\text{low}},r_{\text{high}}\}$.

2.2 Structure-Aware Multi-Stage Diagnostic Process

To model neighborhood context and structure-level dependencies, DiagMIL performs hierarchical feature integration across patch and structure levels.

Attention-based Message Passing (AttnMP). For any graph $G=(V,E)$, each node i updates its representation by aggregating messages from its neighboring nodes set $\mathcal{N}_i$ through an attention-weighted propagation:

$$\text{AttnMP}(H^{(l)},A)=\sigma\big(\tilde{A}\,(H^{(l)}W^{(l)})\big),\quad \tilde{A}_{ij}=\frac{\exp((h_i^{(l)}\cdot h_j^{(l)}))}{\sum_{p\in\mathcal{N}_i}\exp((h_i^{(l)}\cdot h_p^{(l)}))}, \tag{3}$$

where $H^{(l)}\in\mathbb{R}^{|V|\times d_l}$ is the layer-l feature matrix (with $h_i^{(0)}=f_i^{(r)}$), A is the adjacency matrix, and $\tilde{A}$ is the attention-weighted adjacency normalized over

$\mathcal{N}_i$. $W^{(l)}$ is a learnable transformation matrix, and $\sigma(\cdot)$ is a nonlinear activation.
Attention-based Pooling (AttnPool). Within a structure S_k, patch embeddings are aggregated into a structure embedding s_k by attention pooling:

$$\text{AttnPool}(H) = \sum_{i \in V_k} a_i h_i = s_k, \quad a_i = \frac{\exp(w_{\text{pool}}^\top h_i)}{\sum_{p \in V_k} \exp(w_{\text{pool}}^\top h_p)}. \tag{4}$$

This pooling focuses on diagnostically salient regions within each structure.
Low-resolution Coarse Screening (LrCS). When evidence is sufficient at low magnification, pathologists often make an early decision without comparing alternatives. We follow this workflow with low-resolution coarse screening, producing class-wise independent predictions. For each structure $S_k^{(r_{\text{low}})}$, we perform intra-structure aggregation over $\mathcal{E}_k^{(r_{\text{low}})}$ to obtain structure embedding $s_k^{(r_{\text{low}})} = \text{AttnPool}(\text{AttnMP}(\mathbf{f}^{(\mathbf{r}_{\text{low}})}, \mathcal{E}_k^{(r_{\text{low}})}))$. We then build a structure-level graph $G_S^{(r_{\text{low}})} = (\mathbf{s}^{(\mathbf{r}_{\text{low}})}, E_S^{(r_{\text{low}})})$, and apply inter-structure AttnMP$(\cdot)$ to update $s_k^{(r_{\text{low}})}$ and yield $\tilde{s}_k^{(r_{\text{low}})}$. By jointly modeling *intra-structure* patch context and *inter-structure* dependencies, LrCS effectively leverages local feature relationships together with global tissue connectivity. Finally, a readout followed by a sigmoid produces the raw activation, which is L_1-normalized over classes:

$$s^{\text{LrCS}} = \text{Sigmoid}(\text{ReadOut}(G_S^{(r_{\text{low}})})), \quad q_c^{\text{LrCS}} = \frac{s_c^{\text{LrCS}}}{\sum_{j=1}^{C} s_j^{\text{LrCS}}}. \tag{5}$$

High-resolution Fine Diagnosis (HrFD). When LrCS is inconclusive, pathologists zoom in to examine detailed morphology and weigh competing diagnoses to reach a final decision. We follow this workflow with high-resolution fine diagnosis, producing a competitive multi-class prediction. Building on *intra-* and *inter-structure* aggregation in r_{high}, HrFD further incorporates inter-resolution links between corresponding patches. We propagate over the cross-resolution adjacency $\tilde{A}^{(r_{\text{high}} \leftrightarrow r_{\text{low}})}$ and update $H^{(l+1)} = \sigma(\tilde{A}^{(r_{\text{high}} \leftrightarrow r_{\text{low}})} H^{(l)} W^{(l)})$. It enables fine-grained fusion of local contextual cues for the final multi-class prediction:

$$q^{\text{HrFD}} = \text{Softmax}(\text{ReadOut}(G_S^{(r_{\text{high}})})).$$

2.3 Uncertainty-Aware Gating and Cross-Resolution Optimization

To model selective coarse-to-fine diagnosis, we introduce a novel differentiable uncertainty-aware gate driven by LrCS evidence. The gate treats LrCS sigmoid outputs as class-wise diagnostic evidence rather than probability estimates.
For class c, we convert the sigmoid output s_c^{LrCS} into evidence using the odds $e_c = \frac{s_c^{\text{LrCS}}}{1 - s_c^{\text{LrCS}}}$, and parameterize a Dirichlet belief [20]:

$$\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\alpha}), \quad \alpha_c = e_c + 1, \tag{6}$$

where $\alpha_c = 1$ indicates no evidence. The total concentration $A = \sum_c \alpha_c$ captures how much overall evidence LrCS has accumulated for the current slide.

From the expected belief $E[\pi_c] = \alpha_c/A$, we derive the following uncertainty signals. We compute the entropy $\bar{H} = -\frac{1}{\log C}\sum_c E[\pi_c]\log E[\pi_c]$, which reflects how diffusely the belief spreads across classes, and the evidence scarcity $u = \frac{C}{A}$, which captures the overall evidence LrCS accumulates. We also compute the margin $m = E[\pi]_{\text{top1}} - E[\pi]_{\text{top2}}$, the gap between the top two classes, and the positive mass $M = \sum_c s_c^{\text{LrCS}}$, indicating how strongly LrCS supports any class as positive. We combine these signals with an end-to-end trainable linear gate:

$$z = w_H(\bar{H} - 0.5) + w_u u - w_M M - w_m m + b, \quad g = \sigma(z), \tag{7}$$

and mix the two-stage predictions during training as:

$$\hat{q} = (1-g)q^{\text{LrCS}} + gq^{\text{HrFD}}, \tag{8}$$

while at inference we use hard routing, selecting q^{HrFD} if $g > 0.5$ and q^{LrCS} otherwise.

Structure-Level Distillation. To ensure consistency across resolutions, we align structure representations between $r_{\text{high}} \leftrightarrow r_{\text{low}}$. The alignment loss that uses pooled structure embeddings $s_k^{(r)}$ is defined as:

$$\mathcal{L}_{\text{align}} = \text{KL}\left(\text{softmax}(W_t s_k^{(r_{\text{high}})}) \middle\| \text{softmax}(W_s s_k^{(r_{\text{low}})})\right). \tag{9}$$

The training objective is defined as:

$$\mathcal{L} = \text{CE}(\hat{q}, y) + \text{CE}(q^{\text{HrFD}}, y) + \lambda_H \bar{H} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}, \tag{10}$$

where each $\lambda_\cdot$ denotes a hyper-parameter, and the objective combines the final prediction loss with auxiliary terms that supervise HrFD, regularize LrCS uncertainty, and enforce cross-resolution consistency.

3 Experiments

3.1 Implementation Details

Dataset. We leverage a real-world In-house[3] colon cohort and two public breast multi-class WSI datasets. The In-house dataset contains 2,297 colon WSIs labeled into seven classes: Tubular Adenoma, Tubulovillous Adenoma, Traditional Serrated Adenoma, Hyperplastic Polyp, Sessile Serrated Lesion, Inflammatory Polyp, and Lymphoid Polyp. We randomly split the In-house data into a 7:1.5:1.5 ratio, which indicates train, validation, and test, respectively. BRACS [4] is an H&E breast carcinoma WSI dataset for subtyping, where we evaluate its original 7-category setting. BCNB [25] consists of core-needle biopsy WSIs of early breast cancer, where we classify 4-way molecular subtype (i.e., Luminal A-B, HER2(+), and Triple Negative). For the public benchmarks, we follow the official splits.

[3] The approval was granted by the Ethics Review Board SMF-IRB-2024-007 and KH2024-059.

Table 1: Quantitative comparison across three datasets. **Bold** and underline indicate the best and second-best, respectively.

Method	In-house		BCNB [25]		BRACS [4]	
	ACC	AUC	ACC	AUC	ACC	AUC
ABMIL(MR) [15]	$91.51_{0.012}$	$99.20_{0.001}$	$42.84_{0.013}$	$71.89_{0.004}$	$38.88_{0.007}$	$72.51_{0.012}$
TransMIL(MR) [21]	$92.40_{0.013}$	$99.35_{0.001}$	$40.37_{0.041}$	$70.62_{0.017}$	$44.92_{0.018}$	$69.71_{0.035}$
DTFD-MIL(MR) [27]	$92.06_{0.006}$	$99.40_{0.001}$	$44.44_{0.015}$	$73.50_{0.006}$	$43.73_{0.028}$	$68.05_{0.010}$
DSMIL [19]	$92.89_{0.003}$	$99.41_{0.001}$	$\underline{45.35}_{0.002}$	$73.82_{0.019}$	$45.15_{0.056}$	$66.55_{0.041}$
HIPT [7]	$88.87_{0.004}$	$98.50_{0.002}$	$35.43_{0.039}$	$65.16_{0.131}$	$38.17_{0.042}$	$67.79_{0.067}$
ZoomMIL [23]	$91.29_{0.003}$	$99.24_{0.001}$	$43.59_{0.031}$	$72.39_{0.019}$	$45.71_{0.021}$	$72.34_{0.015}$
HAG-MIL [24]	$90.54_{0.010}$	$98.86_{0.002}$	$40.58_{0.013}$	$73.79_{0.018}$	$39.84_{0.158}$	$66.70_{0.056}$
H^2-MIL [14]	$91.73_{0.006}$	$98.85_{0.001}$	$43.05_{0.012}$	$74.23_{0.016}$	$48.78_{0.032}$	$73.21_{0.058}$
HIGT [12]	$\underline{92.97}_{0.007}$	$99.27_{0.002}$	$42.83_{0.023}$	$\underline{74.77}_{0.004}$	$\underline{51.07}_{0.014}$	$76.14_{0.015}$
DAS-MIL [3]	$92.35_{0.008}$	$\underline{99.42}_{0.001}$	$40.84_{0.061}$	$73.83_{0.011}$	$50.08_{0.018}$	$\underline{77.44}_{0.014}$
DiagMIL (Ours)	$\mathbf{93.67}_{0.005}$	$\mathbf{99.44}_{0.001}$	$\mathbf{49.60}_{0.032}$	$\mathbf{78.37}_{0.009}$	$\mathbf{52.92}_{0.021}$	$\mathbf{78.92}_{0.018}$

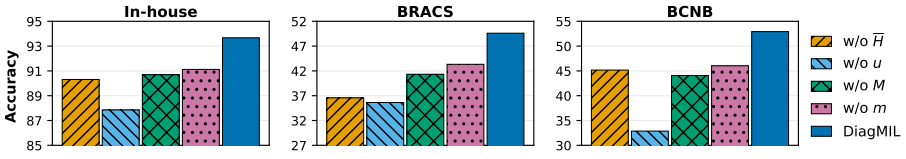

Fig. 2: Gating component study.

Training Settings. For all datasets, we use two resolutions: 10× and 20×, with a patch size of 256. The patches are encoded by a frozen extractor [6], which yields 384-dim embeddings. We train with Adam [16] using a learning rate of 5×10^{-4} and an early stopping patience of 10 epochs. We clamp K to $[3, 300]$ and set $(\lambda, \lambda_H, \lambda_{\text{align}}) = (0.03, 0.05, 0.1)$, running all experiments with three random seeds on an NVIDIA® A6000 GPU.

Comparison Methods. To ensure a fair comparison, we consider MIL baselines that use multi-resolution WSIs as inputs. We include DSMIL [19] as a representative multi-resolution baseline and adopt its fusion scheme to extend ABMIL [15], TransMIL [21] and DTFD-MIL [27] from single- to multi-resolution inputs. We further introduce modern MRMILs that fuse multi-scale information hierarchically, including ZoomMIL [23], HIPT [7] and HAG-MIL [24], and graph-based approaches such as H^2-MIL [14], HIGT [12], and DAS-MIL [3].

3.2 Quantitative Results

Table 1 summarizes the results on three datasets, where DiagMIL achieves the best performance across all metrics. The graph-based methods, which are highlighted in , achieve stronger performance than other MR-MIL approaches by modeling inter-patch relationships and connectivity. This indicates that graph modeling enables a more structured and effective use of multi-resolution information. ZoomMIL and HAG-MIL, which are also motivated to hierarchically model

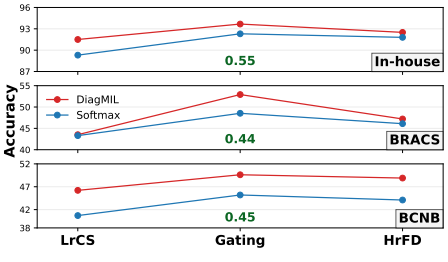

Fig. 3: Stage comparison. Green numbers indicate the HrFD usage rate.

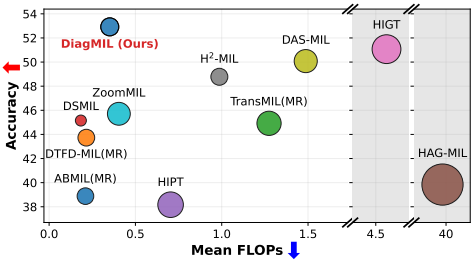

Fig. 4: Cost analysis. Circle size indicates the number of model parameters.

Method	Inference	Preprocess	Total
ABMIL(MR) [15]	2.1	738k	≈738.0k
TransMIL(MR) [21]	5.7	738k	≈738.0k
DTFD-MIL(MR) [27]	2.3	738k	≈738.0k
DSMIL [19]	7.4	738k	≈738.0k
HIPT [7]	4.6	738k	≈738.0k
ZoomMIL [23]	81.8	738k	≈738.1k
HAG-MIL [24]	106.9	738k	≈738.1k
H^2-MIL [14]	68.7	738.2k	≈738.3k
HIGT [12]	38.7	738.9k	≈738.9k
DAS-MIL [3]	10.8	739.0k	≈739.0k
DiagMIL (Ours)	**5.9**	**535.1k**	**≈535.1k**
Δ vs. ABMIL(MR)	+3.8 ms	–	−3.38 min

Table 2: Per-slide processing time (ms).

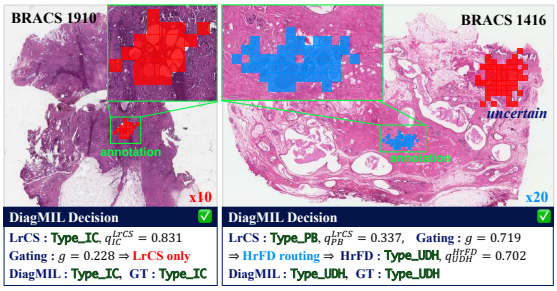

Fig. 5: Qualitative analysis. LrCS, HrFD top-1 attended structure.

the diagnostic process, still use all resolutions for prediction and thus remain limited in performance. In contrast to other MR-MIL methods that fuse information from all resolutions, our model selectively leverages information based on resolution-wise decisions. This selective strategy leads to more effective use of high-resolution evidence and substantial performance gains on challenging datasets such as BCNB and BRACS.

3.3 Gating Analysis

Gating component study. Fig. 2 ablates the four uncertainty signals used by the gate, and the results show that removing any single cue consistently degrades performance. Each cue captures a distinct type of ambiguity in LrCS evidence, and their combination effectively prevents biased routing and reduces mis-escalation. Notably, dropping the evidence-scarcity term u yields the largest performance drop across all datasets. This result underscores the importance of Dirichlet-based evidence modeling for LrCS outputs in effective routing.

Stage comparison. Fig. 3 compares three operating modes: LrCS-only, with gating, and always-on HrFD. Across all datasets, selective gating achieves the best accuracy while invoking high-resolution processing for only 48% of slides on average. In contrast, forcing HrFD on every slide can reduce performance, indicating that high-magnification evidence may introduce nuisance noise that blurs

decision boundaries when coarse evidence is already sufficient. To verify the importance of class-wise independent decisions in LrCS, we replace LrCS's sigmoid prediction with a multi-class softmax head and observe a substantial degradation. Because LrCS makes class-wise independent predictions, it can identify ambiguity that is masked when a multi-class model is forced to choose a single winner. This design enables the gate to detect both cases where multiple classes receive strong support and cases where no class is strongly supported, routing them to HrFD for careful high-resolution diagnosis.

3.4 Further Analysis

Cost and Efficiency Comparison Fig. 4 compares accuracy, mean FLOPs per slide, and model size on BRACS, where DiagMIL achieves the highest accuracy with greater efficiency. Compared to HIGT and DAS-MIL, DiagMIL uses 9.56× and 2.88× fewer parameters, and reduces FLOPs by up to 13.1× compared to HIGT. Concretely, DiagMIL skips HrFD for about 45% of slides, yielding FLOPs comparable to lightweight models. This efficiency arises from our stage-wise design that selectively activates high-resolution computation. It also translates into reduced runtime, as shown in Table 2. While inference time is small and comparable across methods, the end-to-end cost is dominated by preprocessing on gigapixel-scale WSIs. For MR pipelines, this cost further increases because higher resolutions generate many more patches. DiagMIL makes stage-wise predictions, so slides resolved at LrCS skip high-resolution preprocessing. This reduces the total time by 3.38 minutes per slide compared to ABMIL(MR), which is critical when scaling to large samples or real-world clinical deployment.

Qualitative Analysis. We conduct a qualitative analysis to verify that our evidence-driven gating behaves as intended using two BRACS slides with green box annotations (Fig. 5). Since our prediction aggregates structure-level evidence via structure attention pooling, we visualize the top-attended structure to inspect where the model bases its decision. For `BRACS-1910`, the top-1 structure at LrCS (■) overlaps with the annotated region, yielding a confident Type_IC prediction ($q_{\text{IC}}^{\text{LrCS}}$=0.831) with a low gate score g, and thus the model terminates at LrCS without invoking high-resolution analysis. In contrast, for `BRACS-1416`, the top-1 structure at LrCS (■) does not align with the annotated region and the evidence is less confident, resulting in a high g and triggering HrFD. With high-resolution evidence, the model shifts attention (■) to the annotated region and correctly predicts UDH ($q_{\text{UDH}}^{\text{HrFD}}$=0.702). These examples illustrate that our gating escalates computation only when LrCS evidence is insufficient, while structure-level attention provides interpretable, pathology-consistent rationales.

4 Conclusion

We presented DiagMIL, a diagnosis-inspired MR-MIL framework that reframes MR WSI diagnosis as a selective coarse-to-fine decision problem. Motivated by the pathologists' workflow, DiagMIL performs LrCS with class-wise independent

sigmoid evidence and converts it into a Dirichlet belief to derive uncertainty signals for our novel differentiable gate. Based on these signals, DiagMIL makes an early decision at LrCS when coarse evidence is sufficient. Only for inconclusive slides, the gate escalates to HrFD to leverage fine-grained high-resolution evidence. This design avoids unnecessary high-magnification processing while retaining fine-scale modeling capacity when it is diagnostically informative. Across extensive experiments, DiagMIL consistently outperforms strong MR baselines. Our gating studies show that selectively invoking HrFD leads to clearer decision boundaries, substantial cost savings, and improved practical deployability.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) under grant number RS-2022-NR068758. We are also deeply grateful for the generous support provided by the Seegene Medical Foundation in South Korea.

Disclosure of Interests

The authors declare that they have no competing financial or personal interests that could have appeared to influence the work reported in this paper.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence **34**(11), 2274–2282 (2012)
2. Barbosa, D., Ferreira, M., Junior, G.B., Salgado, M., Cunha, A.: Multiple instance learning in medical images: A systematic review. IEEE Access **12**, 78409–78422 (2024)
3. Bontempo, G., Porrello, A., Bolelli, F., Calderara, S., Ficarra, E.: Das-mil: 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)
4. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database **2022**, baac093 (2022)
5. Brunyé, T.T., Mercan, E., Weaver, D.L., Elmore, J.G.: Accuracy is in the eyes of the pathologist: the visual interpretive process and diagnostic accuracy with digital whole slide images. Journal of biomedical informatics **66**, 171–179 (2017)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. CoRR **abs/2104.14294** (2021), `https://arxiv.org/abs/2104.14294`

7. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16144–16155 (2022)
8. Dimitriou, N., Arandjelović, O., Caie, P.D.: Deep learning for whole slide image analysis: an overview. Frontiers in medicine **6**, 264 (2019)
9. Dong, J., Jiang, J., Jiang, K., Li, J., Zhang, Y.: Fast and accurate gigapixel pathological image classification with hierarchical distillation multi-instance learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30818–30828 (2025)
10. Gadermayr, M., Tschuchnig, M.: 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)
11. Ghezloo, F., Wang, P.C., Kerr, K.F., Brunyé, T.T., Drew, T., Chang, O.H., Reisch, L.M., Shapiro, L.G., Elmore, J.G.: An analysis of pathologists' viewing processes as they diagnose whole slide digital images. Journal of Pathology Informatics **13**, 100104 (2022)
12. Guo, Z., Zhao, W., Wang, S., Yu, L.: Higt: Hierarchical interaction graph-transformer for whole slide image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 755–764. Springer (2023)
13. Hong, S., Lee, S., Shin, J., Jeong, J., Yi, M.Y.: Diagnose like a real pathologist: An uncertainty-focused approach for trustworthy multi-resolution multiple instance learning. The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2026)
14. Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., Wang, L.: Hˆ 2-mil: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 933–941 (2022)
15. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Krupinski, E.A., Graham, A.R., Weinstein, R.S.: Characterizing the development of visual search expertise in pathology residents viewing whole slide images. Human pathology **44**(3), 357–364 (2013)
18. Kruskal, J.B.: On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical society **7**(1), 48–50 (1956)
19. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021)
20. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
21. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)

22. Sudin, E., Searjeant, M., Partridge, G., Phillips, P., Hiller, L., Snead, D., Ellis, I., Chen, Y.: Digital pathology: the effect of experience on visual search behavior. Journal of Medical Imaging **9**(3), 035501–035501 (2022)
23. Thandiackal, K., Chen, B., Pati, P., Jaume, G., Williamson, D.F., Gabrani, M., Goksel, O.: Differentiable zooming for multiple instance learning on whole-slide images. In: European Conference on Computer Vision. pp. 699–715. Springer (2022)
24. Xiong, C., Chen, H., Sung, J.J., King, I.: Diagnose like a pathologist: transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. pp. 1587–1595 (2023)
25. Xu, F., Zhu, C., Tang, W., Wang, Y., Zhang, Y., Li, J., Jiang, H., Shi, Z., Liu, J., Jin, M.: Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. Frontiers in Oncology p. 4133 (2021)
26. Yu, X., Luo, H., Hu, J., Zhang, X., Wang, Y., Liang, W., Bei, Y., Song, M., Feng, Z.: Hundredfold accelerating for pathological images diagnosis and prognosis through self-reform critical region focusing. In: IJCAI. pp. 1607–1615 (2024)
27. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18802–18812 (2022)
28. Zhang, Y., Gao, Z., He, K., Li, C., Mao, R.: From patches to wsis: A systematic review of deep multiple instance learning in computational pathology. Information Fusion p. 103027 (2025)