BU-HIM: BAYESIAN UNCERTAINTY-GUIDED HARD INSTANCE MINING FOR WHOLE SLIDE IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Whole slide image (WSI) classification is fundamental to computational pathology, yet its large-scale and weakly annotated nature poses significant challenges for effective representation learning. Existing multiple instance learning (MIL) methods typically rely on deterministic attention mechanisms to identify informative instances, but instances with comparable attention may exhibit substantially different levels of epistemic uncertainty, leading to suboptimal instance selection. To address this issue, we propose BU-HIM, a plug-and-play Bayesian uncertainty-guided hard instance mining framework for weakly supervised WSI classification. Specifically, a Bayesian variational attention mechanism is introduced to jointly characterize instance attention and epistemic uncertainty, where attention variance serves as a complementary uncertainty-aware criterion for identifying informative hard instances. Furthermore, a teacher–student collaborative learning strategy is developed to adaptively optimize the model using the selected hard instances while preserving global contextual information from the entire slide. Extensive experiments on the CAMELYON16 and PANDA benchmark datasets demonstrate that BU-HIM consistently improves the performance of various MIL backbones and outperforms MC-Dropout-based instance selection. These results validate BU-HIM as an effective, uncertainty-aware, and backbone-compatible training framework for weakly supervised WSI classification.

1 INTRODUCTION

Advances in whole-slide scanning technology have enabled pathological specimens to be digitized into gigapixel WSIs, facilitating computational analysis in digital pathology (Madabhushi & Lee, 2016; Komura & Ishikawa, 2018). However, the gigapixel resolution of WSIs makes fine-grained annotation highly labor-intensive, posing challenges for fully supervised learning. To alleviate this limitation, MIL provides an effective weakly supervised framework for WSI analysis (Dietterich et al., 1997; Carbonneau et al., 2018; Campanella et al., 2019). In MIL, each WSI is represented as a bag of patches (instances), while only slide-level labels are required during training. Learning informative slide representations from weakly supervised instances is therefore central to MIL-based WSI classification.

By assigning learnable weights to individual instances during bag aggregation, attention-based MIL selectively aggregates instance-level information to construct slide-level representations (Ilse et al., 2018). Subsequent methods further advanced MIL through improved instance modeling and aggregation strategies (Lu et al., 2021; Li et al., 2021; Shao et al., 2021; Zhang et al., 2022). However, attention is fundamentally designed to quantify the relative contribution of instances to bag-level prediction, providing only a deterministic importance estimate. Such importance alone may not fully characterize an instance: instances with comparable attention can exhibit substantially different uncertainty in their importance estimation. This distinction suggests that importance and estimation uncertainty provide complementary perspectives for instance evaluation. Relying solely on deterministic attention may therefore overlook uncertain yet informative instances that require further optimization.

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.

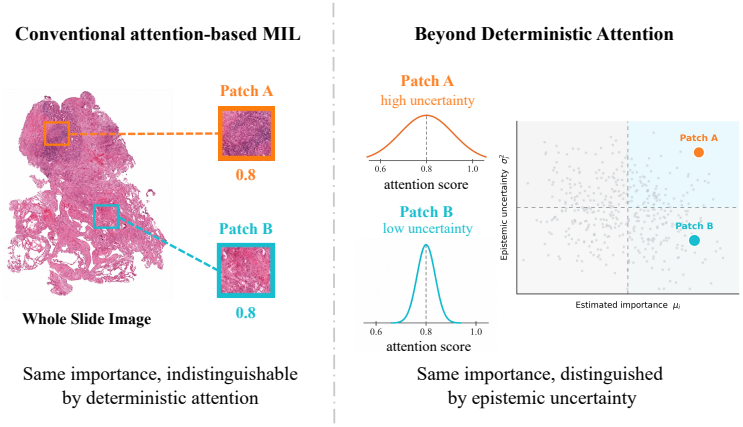

Figure 1: Rethinking instance informativeness in weakly supervised WSI classification.

Motivated by these observations, we propose BU-HIM, a plug-and-play Bayesian uncertainty-guided hard instance mining framework for weakly supervised WSI classification. BU-HIM introduces Bayesian variational attention to characterize each instance through two complementary signals, where attention reflects task relevance and the associated uncertainty captures the model's certainty in the attention estimation. By jointly leveraging these signals, BU-HIM constructs a hard-instance subset by complementarily selecting instances with high estimated importance and uncertain importance estimation. A teacher–student collaborative learning strategy is further developed to balance targeted hard-instance learning and global representation learning, where the Bayesian teacher guides instance evaluation and mining, while the student is jointly optimized on the mined subset and the complete bag.

Extensive experiments on the CAMELYON16 and PANDA benchmark datasets demonstrate that BU-HIM consistently improves multiple representative MIL backbones and outperforms alternative uncertainty estimation strategies. Further analysis confirms that epistemic uncertainty provides complementary information beyond deterministic attention, supporting its effectiveness for uncertainty-aware hard instance mining.

Our main contributions are summarized as follows. **Algorithmic.** We identify and empirically characterize a limitation of deterministic attention-based instance evaluation: a single attention score provides only a point estimate of instance importance and cannot capture its epistemic uncertainty. To address this limitation, we propose BU-HIM, a plug-and-play framework that employs Bayesian variational attention to jointly characterize instance importance and uncertainty in importance estimation. These complementary signals are used to construct a challenging subset containing instances with high task relevance or high uncertainty, which is jointly optimized with the complete bag through teacher–student collaborative learning. **Application.** We systematically evaluate BU-HIM on CAMELYON16 and PANDA across heterogeneous MIL backbones. Comprehensive experiments demonstrate consistent performance improvements and confirm that epistemic uncertainty provides complementary information beyond deterministic attention for instance evaluation and hard instance mining.

2 Related Work

2.1 Multiple Instance Learning for WSI Analysis

Attention-based MIL (ABMIL) (Ilse et al., 2018) has established a foundation for weakly supervised WSI classification, motivating subsequent advances in instance modeling and bag aggregation. CLAM incorporates instance-level clustering to enhance discriminative representation learning (Lu et al., 2021), while DSMIL jointly exploits instance- and bag-level evidence through a dual-stream architecture (Li et al., 2021). Beyond independent instance weighting, TransMIL models inter-instance dependencies using Transformer-based aggregation (Shao et al., 2021), whereas DTFD-MIL improves weakly supervised learning through pseudo-bag construction and double-tier

feature distillation (Zhang et al., 2022). Recent studies have further investigated confounding-aware bag representations and instance attributes (Lin et al., 2023; Cai et al., 2025). Beyond aggregation, another line of research investigates which instances should be emphasized during training.

2.2 Hard Instance Mining in MIL

Hard instances are particularly valuable for MIL optimization because they expose aspects of the data that remain insufficiently learned by the current model, providing informative training signals beyond instances that have already been well modeled (Shrivastava et al., 2016). MHIM-MIL directly explores hard instance learning for WSI classification through masked hard instance mining, where a momentum teacher guides attention-based masking to encourage the student to learn from challenging instances rather than repeatedly relying on highly activated regions (Tang et al., 2023). Closely related instance-selection strategies regulate instance participation from different perspectives. DPIS integrates dynamic instance sampling with reinforcement learning to adaptively select instances during MIL optimization (Zheng et al., 2024), while GMMamba employs attention-based group masking to filter redundant and uninformative instances and improve instance-aware representation learning (Zheng et al., 2025). Curriculum-based MIL further adjusts training difficulty to reduce bias during histopathology classification (Mi et al., 2025). Together, these studies highlight the broader value of selectively regulating instance participation during MIL optimization.

However, the effectiveness of such strategies critically depends on how instances are evaluated for selection. Existing approaches predominantly derive selection or masking decisions from deterministic model responses, such as attention, feature similarity, or learned sampling scores, providing limited information about the model's certainty in these evaluations. This motivates a richer characterization of instances for hard instance mining beyond deterministic criteria alone.

2.3 Uncertainty Modeling in MIL

Uncertainty estimation provides a principled means of characterizing the reliability of model predictions and separating uncertainty arising from data from that associated with model parameters (Kendall & Gal, 2017; Abdar et al., 2021; Gawlikowski et al., 2023). Recent studies have introduced probabilistic and uncertainty-aware modeling into MIL from different perspectives, including Gaussian-process-based probabilistic attention and reliability-oriented uncertainty modeling (Schmidt et al., 2024; de Vente et al., 2024). Bayes-MIL formulates attention probabilistically and derives patch-level uncertainty as an interpretability measure for identifying regions of interest (Cui et al., 2023). E^2-MIL introduces an evidential instance classifier to estimate instance-level predictive uncertainty, improving prediction reliability and further using uncertainty to filter instances with noisy pseudo-labels during training (Shi et al., 2024). At the bag level, RPMIL replaces deterministic bag embeddings with probabilistic feature distributions to capture uncertainty in slide representations (Zhao et al., 2025), while SIB-MIL introduces Bayesian modeling to improve uncertainty quantification and robustness under weak supervision (Chen et al., 2026). These studies demonstrate the utility of uncertainty modeling across different levels of MIL, ranging from instance interpretation and reliability estimation to probabilistic bag representation.

Despite these advances, existing uncertainty-aware MIL methods mainly focus on uncertainty quantification, interpretability, or robustness, while leveraging epistemic uncertainty as a complementary signal to deterministic attention for hard instance mining remains underexplored. In contrast, BU-HIM incorporates epistemic uncertainty into hard instance evaluation, allowing uncertainty in instance importance estimation to directly guide hard-instance selection and targeted optimization.

3 Methodology

We propose **BU-HIM**, a Bayesian uncertainty-guided hard instance mining framework for weakly supervised WSI classification. BU-HIM comprises uncertainty-aware instance evaluation and dual-view teacher–student optimization. Through Bayesian variational attention, each instance is characterized by two complementary signals: its estimated importance and the epistemic uncertainty associated with this estimation. These signals are used to construct a hard-instance subset from importance-aware and uncertainty-aware perspectives. The student subsequently learns from both

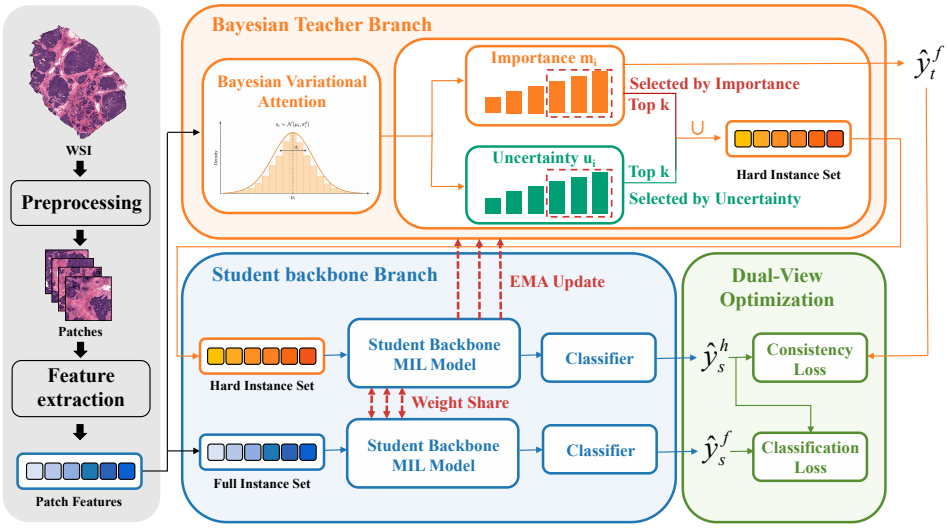

Figure 2: Overview of the proposed BU-HIM framework.

the mined hard-instance view and the full-bag view, while the Bayesian teacher guides instance evaluation and provides a full-bag prediction reference for learning from the selected instances.

3.1 Problem Formulation and Motivation

In MIL-based WSI classification, each WSI is represented as a bag of image patches. Given a WSI, its patch bag is denoted as $\mathcal{X} = \{x_i\}_{i=1}^{N}$, where x_i represents the i-th instance and N denotes the number of instances. During training, only the bag-level label $Y \in \mathcal{C}$ is available, while instance-level annotations remain unknown. After feature extraction, the WSI is represented as an instance feature bag $\mathcal{Z} = \{z_i\}_{i=1}^{N}$, where $z_i \in \mathbb{R}^D$ denotes the feature representation of the i-th instance.

Attention-based MIL learns an instance-level weighting function and aggregates the instance features into a bag representation $F = \sum_{i=1}^{N} a_i z_i$. Following the gated attention mechanism (Ilse et al., 2018), the normalized attention weight is computed as

$$a_i = \frac{\exp\left(w^\top \left[\tanh(Vz_i) \odot \sigma(Uz_i)\right]\right)}{\sum_{j=1}^{N} \exp\left(w^\top \left[\tanh(Vz_j) \odot \sigma(Uz_j)\right]\right)}, \tag{1}$$

where w, V, and U are learnable parameters, $\sigma(\cdot)$ denotes the sigmoid function, and $\odot$ represents element-wise multiplication. A classifier $C(\cdot)$ then produces the slide-level prediction $\hat{Y} = C(F)$.

Although attention-based MIL has demonstrated strong capability in identifying discriminative instances, the attention score itself only provides a deterministic estimation of instance importance. Specifically, a_i reflects how much an instance contributes to the current slide-level prediction, but does not characterize the reliability of this importance estimation. Consequently, instances with similar attention values may correspond to substantially different levels of epistemic uncertainty. Some instances may be consistently recognized as important by the model, while others may receive unstable importance estimates due to ambiguous evidence or insufficient optimization.

Hard instance mining aims to identify instances that provide additional optimization signals for improving model learning. Therefore, the quality of instance evaluation plays a crucial role in determining which instances should receive further emphasis during training. Existing strategies typically rely on deterministic instance scores, such as attention magnitude, to measure instance importance for selection. However, these scores only describe the estimated contribution of instances and overlook the uncertainty associated with such estimation. Motivated by this observation, BU-HIM introduces epistemic uncertainty as a complementary signal to attention importance, enabling a more comprehensive instance evaluation for uncertainty-guided hard instance mining.

3.2 Bayesian Variational Attention

To explicitly model uncertainty in instance importance estimation, we introduce Bayesian variational attention. Unlike conventional attention mechanisms with deterministic parameters, the weights ω of the Bayesian attention module are characterized by a variational posterior $q_\psi(\omega)$ (Blundell et al., 2015). This formulation represents uncertainty over model parameters and enables epistemic uncertainty in the resulting importance estimates to be quantified (Kendall & Gal, 2017).

Given an instance feature z_i, the Bayesian attention parameterizes a latent attention response as

$$[\mu_i, \rho_i] = g_B(z_i; \omega), \qquad \omega \sim q_\psi(\omega), \tag{2}$$

where μ_i and ρ_i denote the mean and predicted log-variance of the latent attention response, respectively. The corresponding variance is $v_i = \exp(\rho_i)$, from which the latent response is modeled as $r_i \sim \mathcal{N}(\mu_i, v_i)$. We define the instance importance score as $m_i = \sigma(\mu_i)$. Here, m_i is an unnormalized score used for instance evaluation and selection, whereas a_i in Section 3.1 denotes the normalized attention weight used for bag aggregation.

Since v_i characterizes uncertainty before the nonlinear transformation, we use a first-order Delta approximation to estimate the variance of the transformed importance response:

$$u_i \approx \mathrm{Var}[\sigma(r_i)] \approx [\sigma'(\mu_i)]^2 v_i = [m_i(1 - m_i)]^2 v_i. \tag{3}$$

The resulting u_i quantifies the epistemic uncertainty associated with the importance estimation. A larger u_i indicates that the estimated importance is more sensitive to variations in the Bayesian attention parameters, whereas a smaller value indicates a more stable estimate. Each instance is therefore characterized by the complementary pair (m_i, u_i), where m_i measures estimated importance and u_i measures uncertainty in that estimate.

The variational posterior is regularized toward its prior through $\mathcal{L}_{\mathrm{var}} = D_{\mathrm{KL}}(q_\psi(\omega) \,\|\, p(\omega))$, which constrains the Bayesian attention parameters and prevents unconstrained uncertainty estimation.

3.3 Uncertainty-Guided Hard Instance Mining

Given the complementary importance–uncertainty pair (m_i, u_i), we construct a hard-instance view for targeted optimization. For an instance feature bag $\mathcal{Z} = \{z_i\}_{i=1}^N$ and a selection ratio β, the number of instances selected under each criterion is defined as $k = \max(1, \lfloor \beta N \rfloor)$. The instances are independently ranked according to their importance and uncertainty scores, and the hard-instance view is constructed as

$$\begin{aligned} \mathcal{I}_m &= \mathrm{TopK}(\mathbf{m}, k), & \mathcal{I}_u &= \mathrm{TopK}(\mathbf{u}, k), \\ \mathcal{I}_h &= \mathcal{I}_m \cup \mathcal{I}_u, & \mathcal{Z}^h &= \{z_i \mid i \in \mathcal{I}_h\}, \end{aligned} \tag{4}$$

where $\mathbf{m} = [m_1, \ldots, m_N]$ and $\mathbf{u} = [u_1, \ldots, u_N]$ denote the instance-level importance and uncertainty sequences, respectively. Importance-aware selection preserves discriminative evidence associated with the current bag-level prediction, while uncertainty-aware selection introduces instances whose importance estimates remain unstable and may therefore be insufficiently modeled.

This set-level combination does not require every retained instance to be simultaneously important and uncertain. Instead, it constructs a complementary hard-instance view that preserves currently identified discriminative evidence while extending the learning focus toward insufficiently modeled instances.

3.4 Dual-View Teacher–Student Optimization

Learning exclusively from the hard-instance view may discard contextual information contained in the complete WSI. We therefore optimize the same student model f_θ using both views, producing the hard-instance prediction $\hat{\mathbf{y}}_s^h = f_\theta(\mathcal{Z}^h)$ and the full-bag prediction $\hat{\mathbf{y}}_s^f = f_\theta(\mathcal{Z})$. The former emphasizes instances identified by the proposed mining strategy, while the latter preserves the complete slide context. Both predictions are supervised by the bag-level label Y:

$$\mathcal{L}_{\mathrm{cls}} = \lambda_h \,\mathrm{CE}\left(\hat{\mathbf{y}}_s^h, Y\right) + (1 - \lambda_h)\,\mathrm{CE}\left(\hat{\mathbf{y}}_s^f, Y\right), \tag{5}$$

where λ_h controls the contribution of the hard-instance branch.

Meanwhile, the Bayesian teacher g_ϕ processes the complete bag and produces the full-bag prediction $\hat{\mathbf{y}}_t^f = g_\phi(\mathcal{Z})$. To reduce the prediction deviation introduced by instance selection, the teacher prediction serves as an auxiliary reference for the hard-instance student prediction:

$$\mathcal{L}_{\text{con}} = \left\| \text{softmax}\left(\hat{\mathbf{y}}_s^h\right) - \text{sg}\left[\text{softmax}\left(\hat{\mathbf{y}}_t^f\right)\right] \right\|_2^2, \tag{6}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}}. \tag{7}$$

Since the Bayesian teacher and different student backbones are not fully identical in architecture, only their structurally compatible parameters are initialized and updated through exponential moving average, following the general principle of weight-averaged teacher models (Tarvainen & Valpola, 2017). Denoting the compatible parameter set by Ω, the teacher parameters are updated as $\phi_j \leftarrow \alpha\phi_j + (1-\alpha)\theta_j$ for $j \in \Omega$, where α is the momentum coefficient.

During the first five epochs, instance mining is disabled to warm up the student on complete bags. After warm-up, the hard-instance view is dynamically reconstructed at each training iteration. At inference time, only the student operating on the complete bag is retained, introducing no additional inference overhead.

4 EXPERIMENTS

We evaluate BU-HIM on CAMELYON16 and PANDA using three repetitions of three-fold cross-validation. We compare with representative MIL baselines including Mean-pooling, Max-pooling, ABMIL, DSMIL, CLAM-MB, TransMIL, MHIM-MIL, RRT-MIL, and FR-MIL. BU-HIM is integrated with ABMIL, DSMIL, and RRT-MIL for backbone evaluation. All experiments use ResNet-50 extracted features and Adam optimization. Details are provided in Appendix.

Table 1: Quantitative comparison with existing MIL methods on CAMELYON16 and PANDA.

Method	Accuracy (%)	AUC (%)	F1-score (%)	Precision (%)	Recall (%)
CAMELYON16					
Mean-pooling	76.76 ± 2.26	79.87 ± 1.81	66.12 ± 2.78	69.19 ± 5.92	64.03 ± 5.98
Max-pooling	85.23 ± 7.45	87.41 ± 8.44	78.87 ± 9.79	81.39 ± 11.87	76.77 ± 8.86
ABMIL	**89.25 ± 1.59**	90.75 ± 1.89	**83.95 ± 2.41**	**89.54 ± 4.13**	79.24 ± 4.31
DSMIL	88.51 ± 2.04	90.48 ± 1.33	83.05 ± 2.88	87.56 ± 5.06	79.24 ± 4.42
CLAM-MB	88.62 ± 2.13	90.59 ± 1.82	83.38 ± 2.82	86.97 ± 4.77	80.17 ± 2.28
TransMIL	88.03 ± 1.84	90.92 ± 1.84	82.11 ± 2.65	87.67 ± 4.23	77.35 ± 3.33
MHIM-MIL	88.25 ± 1.43	91.16 ± 1.26	82.67 ± 1.87	87.07 ± 3.91	78.81 ± 2.47
RRT-MIL	88.25 ± 1.49	91.44 ± 1.74	82.89 ± 2.29	85.85 ± 2.89	80.28 ± 4.18
FR-MIL	87.66 ± 2.45	90.72 ± 2.15	82.31 ± 2.89	84.69 ± 6.78	80.49 ± 3.89
ABMIL + BU-HIM	88.90 ± 1.14	91.45 ± 0.65	83.60 ± 1.22	88.47 ± 5.00	79.54 ± 3.39
DSMIL + BU-HIM	88.83 ± 0.81	91.18 ± 1.10	83.27 ± 1.22	89.04 ± 2.69	78.29 ± 2.45
RRT-MIL + BU-HIM	88.34 ± 1.22	**91.76 ± 0.73**	83.26 ± 1.56	85.16 ± 3.86	**81.63 ± 3.16**
PANDA					
Mean-pooling	61.76 ± 0.80	88.60 ± 0.32	55.75 ± 0.79	57.96 ± 0.68	55.13 ± 0.65
Max-pooling	59.29 ± 1.24	86.72 ± 0.35	51.09 ± 1.18	52.10 ± 0.90	51.46 ± 0.77
ABMIL	65.71 ± 0.91	90.36 ± 0.30	59.48 ± 0.80	60.67 ± 0.75	59.23 ± 0.80
DSMIL	63.82 ± 0.94	89.69 ± 0.35	56.47 ± 1.50	58.24 ± 0.86	56.67 ± 1.09
CLAM-MB	65.61 ± 0.39	90.27 ± 0.31	59.61 ± 0.45	61.39 ± 0.75	59.19 ± 0.54
TransMIL	64.84 ± 0.84	89.82 ± 0.21	58.24 ± 0.94	59.65 ± 1.10	58.25 ± 0.88
MHIM-MIL	63.57 ± 1.08	89.60 ± 0.44	56.32 ± 1.37	58.98 ± 0.67	56.76 ± 1.20
RRT-MIL	66.72 ± 0.99	90.93 ± 0.49	60.61 ± 0.85	62.01 ± 0.73	60.34 ± 0.75
ABMIL + BU-HIM	65.60 ± 0.78	90.39 ± 0.19	59.13 ± 0.98	60.37 ± 0.68	58.99 ± 1.04
DSMIL + BU-HIM	66.38 ± 0.79	90.73 ± 0.24	60.30 ± 1.19	61.43 ± 0.84	60.05 ± 1.17
RRT-MIL + BU-HIM	**67.67 ± 0.61**	**91.36 ± 0.30**	**61.93 ± 0.75**	**62.86 ± 0.66**	**61.56 ± 0.84**

4.1 Comparison with Existing MIL Methods

To evaluate the classification performance and architectural compatibility of BU-HIM, we compare it with representative MIL methods on CAMELYON16 and PANDA. In addition to evaluating standalone MIL baselines, we integrate BU-HIM with ABMIL, DSMIL, and RRT-MIL, allowing direct comparison of each backbone before and after introducing the proposed framework. The quantitative results are reported in Table 1.

On CAMELYON16, BU-HIM consistently improves AUC across all three student backbones, including ABMIL, DSMIL, and RRT-MIL. This demonstrates that the proposed uncertainty-guided instance evaluation is compatible with different MIL aggregation mechanisms. Moreover, BU-HIM reduces AUC variance across runs, suggesting more stable instance evaluation under different data partitions.

On PANDA, BU-HIM achieves larger improvements than on CAMELYON16, particularly for DSMIL and RRT-MIL. We attribute this improvement to the challenging multi-class setting, where subtle morphological differences and intra-class heterogeneity make deterministic attention-based instance evaluation less reliable. By incorporating uncertainty in importance estimation, BU-HIM provides additional signals for identifying instances that require further optimization.

Overall, BU-HIM improves performance across heterogeneous MIL backbones and datasets, indicating that uncertainty-guided hard instance mining provides a complementary optimization strategy beyond deterministic attention-based selection.

4.2 Comparison of Different Instance Mining Strategies

To examine whether the performance gain of BU-HIM arises from uncertainty-guided instance mining rather than simply reducing the number of training instances, we conduct an independent controlled comparison using ABMIL on CAMELYON16. The compared strategies include the original ABMIL without instance mining, random instance selection, MC-Dropout-based uncertainty selection, and the complete BU-HIM framework. Random selection retains the same number of instances as BU-HIM without using importance or uncertainty information. MC-Dropout estimates instance uncertainty through repeated stochastic forward passes (Gal & Ghahramani, 2016) and selects instances according to the resulting uncertainty scores. Within this comparison, all methods use the same patch features, data partitions, training configurations, and evaluation protocol. The results are reported in Table 2.

Table 2: Comparison of different instance mining strategies on CAMELYON16.

Method	AUC (%)
ABMIL	91.03 ± 1.49
ABMIL + Random Selection	91.09 ± 1.53
ABMIL + MC-Dropout Selection	$\underline{91.34 \pm 0.06}$
ABMIL + BU-HIM	$\mathbf{91.57 \pm 1.09}$

Random instance selection only increases the AUC of ABMIL from 91.03% to 91.09%, while increasing the standard deviation from 1.49 to 1.53. This result indicates that merely reducing the number of training instances is insufficient to produce a consistent performance improvement. In comparison, MC-Dropout-based selection improves AUC to 91.34%, suggesting that model uncertainty provides a more informative criterion for instance selection than random subsampling.

BU-HIM achieves the highest AUC of 91.57%, outperforming the original ABMIL, random selection, and MC-Dropout-based selection by 0.54, 0.48, and 0.23 percentage points, respectively. Unlike random or uncertainty-only selection, BU-HIM uses Bayesian variational attention to jointly characterize instance importance and uncertainty, and combines their candidate sets through a set union to preserve importance-aware evidence while incorporating uncertain instances. Its advantage over random selection indicates that the improvement cannot be explained solely by instance subsampling. Its performance over the evaluated MC-Dropout alternative further suggests that complementary selection based on importance and Bayesian uncertainty provides a more effective training view.

4.3 Decision Sensitivity Beyond Deterministic Attention

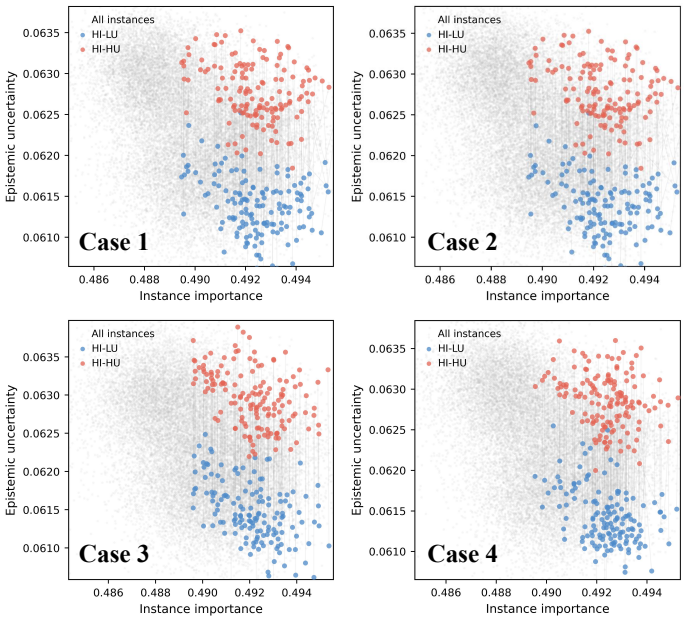

Figure 3: Instance selection in the importance–uncertainty space for four representative cases.

To examine whether epistemic uncertainty provides complementary information beyond deterministic attention, we construct attention-matched instance pairs with different uncertainty levels. For each WSI, the top 20% of instances by attention are first selected as high-attention candidates. Within this set, the top and bottom 20% by uncertainty form the high-attention high-uncertainty (HI-HU) and high-attention low-uncertainty (HI-LU) groups, respectively. Each HI-HU instance is then paired with the HI-LU instance having the closest attention score. As shown in Figure 3, the matched groups exhibit similar attention scores but remain clearly separated in uncertainty.

To assess their influence on bag-level decisions, we separately remove matched HI-HU or HI-LU instances using removal budgets of 50, 100, and 200. Decision sensitivity is quantified by the absolute changes in prediction margin, true-class probability, and classification loss. Attention matching helps control for differences in deterministic attention, allowing the comparison to focus on the additional information captured by uncertainty.

Table 3: Decision sensitivity of attention-matched HI-HU and HI-LU instances.

Budget	Group	$\lvert\Delta\text{Margin}\rvert$	$\lvert\Delta P_{\text{true}}\rvert$	$\lvert\Delta\mathcal{L}\rvert$
50	HI-HU	0.003331 ± 0.016894	0.000518 ± 0.002279	0.000780 ± 0.002922
50	HI-LU	0.000262 ± 0.002366	0.000038 ± 0.000331	0.000091 ± 0.001183
100	HI-HU	0.004295 ± 0.017955	0.000689 ± 0.002593	0.001026 ± 0.003422
100	HI-LU	0.000289 ± 0.002376	0.000043 ± 0.000332	0.000097 ± 0.001184
200	HI-HU	0.005110 ± 0.019333	0.000832 ± 0.002925	0.001242 ± 0.003947
200	HI-LU	0.000333 ± 0.002432	0.000050 ± 0.000345	0.000107 ± 0.001190

As shown in Table 3, removing HI-HU instances consistently induces larger prediction changes than removing attention-matched HI-LU instances across all budgets and metrics. At a removal budget of 50, the changes in prediction margin, true-class probability, and classification loss are approximately $12.71\times$, $13.63\times$, and $8.57\times$ larger for HI-HU instances, respectively, and the same trend persists at larger budgets. These results show that instances with similar deterministic attention

can exhibit substantially different decision sensitivity. Epistemic uncertainty therefore provides a complementary signal for instance evaluation beyond attention alone.

4.4 Qualitative Analysis of Attention Localization

To qualitatively examine how BU-HIM affects spatial attention allocation, we compare ABMIL with ABMIL+BU-HIM on CAMELYON16. Representative WSIs with tumor-region annotations are selected, and attention maps are visualized using the same color range for fair comparison. Additional feature representation analysis is provided in Appendix B.1.

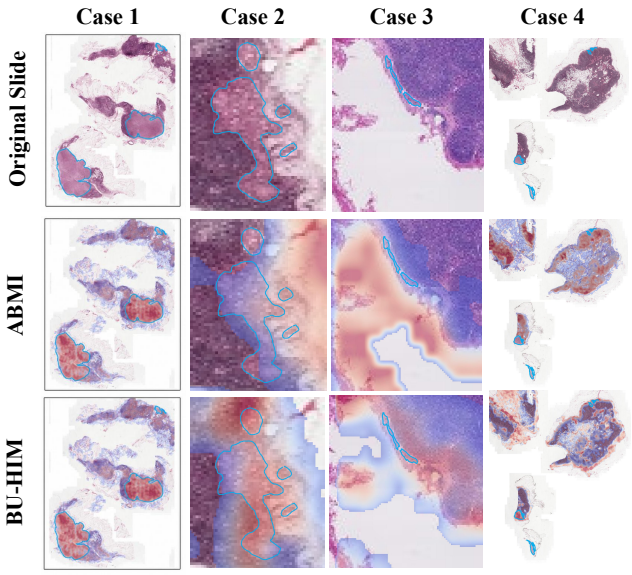

Figure 4: Qualitative comparison of attention heatmaps produced by ABMIL and BU-HIM across four representative cases.

As shown in Figure 4, both ABMIL and ABMIL+BU-HIM identify tumor-related regions with high attention responses. However, ABMIL exhibits relatively diffuse attention patterns, including isolated responses outside tumor boundaries. In comparison, BU-HIM produces more concentrated attention responses that better correspond to tumor annotations, particularly for small or spatially dispersed lesions. These observations suggest that uncertainty-guided instance evaluation helps refine attention allocation beyond deterministic attention estimation. The improvement is especially evident in challenging regions where discriminative evidence is sparse or spatially fragmented. This further indicates that incorporating epistemic uncertainty can help the model focus more consistently on pathologically relevant instances.

5 Conclusion

In this work, we presented BU-HIM, a Bayesian uncertainty-guided hard instance mining framework for weakly supervised WSI classification. BU-HIM models instance importance together with epistemic uncertainty and leverages these complementary signals for uncertainty-aware instance selection. A dual-view teacher–student strategy further enables targeted hard-instance optimization while preserving complete slide context. Experiments on CAMELYON16 and PANDA demonstrate consistent improvements across multiple MIL backbones, indicating the effectiveness and backbone compatibility of the proposed framework. Further analyses show that instances with comparable deterministic attention can exhibit substantially different decision sensitivity, highlighting the complementary information provided by epistemic uncertainty. Qualitative results further suggest that BU-HIM encourages more discriminative instance representations and more concentrated attention localization. Overall, these findings support uncertainty-aware instance evaluation as a promising direction for improving hard instance mining in weakly supervised WSI classification.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and language polishing. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This study uses the publicly released CAMELYON16 and PANDA benchmark datasets solely for methodological research. No new patient data were collected, no human participants were directly involved, and no clinical decisions were made as part of this work. BU-HIM is presented as a research framework rather than a clinically validated diagnostic system. Any potential clinical deployment would require additional external validation and appropriate ethical and regulatory review.

REPRODUCIBILITY STATEMENT

The formulation of BU-HIM, including Bayesian variational attention, uncertainty-guided instance selection, the dual-view training objective, and the teacher update strategy, is described in Section 3. Appendix A provides dataset descriptions, evaluation protocols, implementation details, and principal hyperparameters. All comparative experiments use consistent feature representations, data partitions, and evaluation protocols, and the reported results are aggregated over three repetitions of three-fold cross-validation.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. doi: 10.1016/j.inffus.2021.05.008.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1613–1622. PMLR, 2015.

Linghan Cai, Shenjin Huang, Ye Zhang, Jinpeng Lu, and Yongbing Zhang. AttriMIL: Revisiting attention-based multiple instance learning for whole-slide pathological image classification from a perspective of instance attributes. *Medical Image Analysis*, 103:103631, 2025.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Angela Miraflor, Vitor W. K. 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. doi: 10.1038/s41591-019-0508-1.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77: 329–353, 2018. doi: 10.1016/j.patcog.2017.10.009.

Yihang Chen, Tsai Hor Chan, Jianning Chen, Li Liang, Guosheng Yin, and Lequan Yu. SIB-MIL: Sparsity-induced bayesian neural network for robust multiple instance learning on whole slide image analysis. *IEEE Transactions on Medical Imaging*, 45(4):1638–1650, 2026. doi: 10.1109/TMI.2025.3638243.

Yufei Cui, Ziquan Liu, Xiangyu Liu, Xue Liu, Cong Wang, Tei-Wei Kuo, Chun Jason Xue, and Antoni B. Chan. Bayes-MIL: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In *International Conference on Learning Representations*, 2023.

Coen de Vente, Bram van Ginneken, Carel B. Hoyng, Caroline C. W. Klaver, and Clara I. Sánchez. Uncertainty-aware multiple-instance learning for reliable classification: Application to optical coherence tomography. *Medical Image Analysis*, 97:103259, 2024. doi: 10.1016/j.media.2024.103259.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71, 1997. doi: 10.1016/S0004-3702(96)00034-3.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059. PMLR, 2016.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023. doi: 10.1007/s10462-023-10562-9.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 2018.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018. doi: 10.1016/j.csbj.2018.01.001.

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*, pp. 14318–14328, 2021.

Tiancheng Lin, Zhimiao Yu, Hongming Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19830–19839, 2023.

Ming Y. Lu, Drew F. K. 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*, 5:555–570, 2021. doi: 10.1038/s41551-020-00682-w.

Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016. doi: 10.1016/j.media.2016.06.037.

Zihao Mi, Jianan Zhang, Xueyu Liu, Guanghui Yue, Junhong Yue, Mingqiang Wei, Yidi Li, and Yongfei Wu. Multi-instance curriculum learning for histopathology image classification with bias reduction. *Medical Image Analysis*, 105:103647, 2025.

Arne Schmidt, Pablo Morales-Álvarez, and Rafael Molina. Probabilistic attention based on gaussian processes for deep multiple instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10909–10922, 2024. doi: 10.1109/TNNLS.2023.3245329.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2136–2147, 2021.

Jiangbo Shi, Chen Li, Tieliang Gong, and Huazhu Fu. E^2-mil: An explainable and evidential multiple instance learning framework for whole slide image classification. *Medical Image Analysis*, 97:103294, 2024. doi: 10.1016/j.media.2024.103294.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 761–769, 2016.

Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4078–4087, 2023.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yutong 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*, pp. 18802–18812, 2022.

Zhikang Zhao, Kaitao Chen, and Jing Zhao. RPMIL: Rethinking uncertainty-aware probabilistic multiple instance learning for whole slide pathology diagnosis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2467–2475, 2025. doi: 10.24963/ijcai.2025/275.

Tingting Zheng, Kui Jiang, and Hongxun Yao. Dynamic policy-driven adaptive multi-instance learning for whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8028–8037, 2024.

Tingting Zheng, Hongxun Yao, Kui Jiang, Yaming Xiao, and Shanshan Zhao. GMMamba: Group masking mamba for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9935–9944, 2025.

A Experimental Details

A.1 Dataset Details

We evaluate BU-HIM on two representative whole-slide image (WSI) classification benchmarks with different diagnostic characteristics. CAMELYON16 focuses on metastatic tumor detection in lymph-node sections, where discriminative tumor regions usually occupy only a small portion of the whole slide. This sparse evidence setting provides a challenging benchmark for evaluating the ability of MIL methods to identify informative instances.

PANDA targets prostate cancer grading from histopathological images. Compared with binary metastasis detection, PANDA contains multiple morphological categories with subtle inter-class differences and substantial intra-class heterogeneity, providing a challenging multi-class setting for WSI classification.

The statistics of the datasets are summarized in Table 4.

Table 4: Dataset details used in our experiments.

Dataset	Task	Slides	Classes
CAMELYON16	Metastasis detection	400	2
PANDA	Prostate grading	10,616	6

A.2 Evaluation Protocol

Following previous MIL studies, we perform three repetitions of three-fold cross-validation using predefined data splits and report the mean and standard deviation over all runs.

Classification performance is evaluated using Accuracy, AUC, F1-score, Precision, and Recall. AUC is treated as the primary metric because it evaluates the ranking capability of the model without

relying on a specific classification threshold. For the multi-class PANDA dataset, AUC is computed using macro-averaged one-vs-rest evaluation, while F1-score, Precision, and Recall are macro-averaged across classes.

A.3 Baseline and Backbone Details

We compare BU-HIM with representative MIL methods covering different instance aggregation and representation modeling strategies. For all methods, the same pre-extracted ResNet-50 patch features and predefined data splits are used to ensure a consistent evaluation setting.

Mean-pooling directly averages all instance features within a WSI to obtain the bag-level representation, while **Max-pooling** performs element-wise maximum aggregation over all instances. These two methods serve as simple non-attention-based MIL baselines.

Mean-pooling directly averages all instance features within a WSI to obtain the bag-level representation, while **Max-pooling** performs element-wise maximum aggregation over all instances. These two methods serve as simple non-attention-based MIL baselines.

ABMIL (Ilse et al., 2018) employs gated attention to assign learnable importance weights to instances and aggregate them into a weighted bag representation. **DSMIL** (Li et al., 2021) adopts a dual-stream architecture that jointly exploits instance-level and bag-level evidence. **CLAM-MB** (Lu et al., 2021) augments attention-based aggregation with instance-level clustering and multiple attention branches, while **TransMIL** (Shao et al., 2021) models inter-instance dependencies using Transformer-based aggregation.

MHIM-MIL (Tang et al., 2023) introduces masked hard instance mining with a momentum teacher to encourage learning from challenging instances rather than repeatedly relying on highly activated regions. **RRT-MIL** models inter-instance relations through relation-aware representation learning and is used as one of the student backbones for evaluating the compatibility of BU-HIM with structured MIL architectures. **FR-MIL** is additionally evaluated on CAMELYON16 following its binary WSI classification setting.

To evaluate the backbone compatibility of BU-HIM, we integrate the proposed framework with ABMIL, DSMIL, and RRT-MIL. The original aggregation and classification architecture of each backbone is retained, while BU-HIM introduces Bayesian teacher guidance, uncertainty-guided hard-instance construction, and dual-view optimization during training. At inference time, only the student backbone operating on the full instance bag is retained.

Random Selection randomly samples the same number of instances as the BU-HIM hard-instance view for each WSI, serving to test whether the performance gain can be attributed merely to instance subsampling.

MC-Dropout Selection estimates instance uncertainty through repeated stochastic forward passes with dropout enabled, following the Bayesian interpretation of dropout (Gal & Ghahramani, 2016). Instances with higher estimated uncertainty are retained for training. This baseline compares BU-HIM with a conventional sampling-based uncertainty estimation strategy under the same feature representation, data partition, and training protocol.

A.4 Implementation Details

All experiments are implemented in PyTorch. Following common MIL settings, each WSI is processed as an individual bag. Patch features are extracted using ResNet-50 and represented as 1024-dimensional vectors.

All models are optimized using Adam with an initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-5}. Cosine learning-rate annealing is adopted during training. Models are trained for up to 100 epochs using cross-entropy supervision. Early stopping is applied with a patience of 30 epochs for CAMELYON16 and 20 epochs for PANDA. The checkpoint achieving the highest validation AUC is selected for testing.

BU-HIM is instantiated on three representative student backbones: ABMIL, DSMIL, and RRT-MIL. The detailed backbone descriptions are provided in Appendix A.3.The Bayesian teacher is initialized from compatible student parameters and updated using exponential moving average with

momentum 0.99. The first five epochs are used for warm-up without instance mining. After warm-up, the hard-instance subset is dynamically constructed during training. The selection ratio β is set to 0.7. The hard-instance classification loss weight, consistency loss weight, and variational regularization weight are set to 0.7, 1.0, and 1.0, respectively. The dropout rate is set to 0.25.

B Additional Qualitative Results

B.1 Feature Representation Visualization

To further investigate how BU-HIM affects instance-level feature representations, we visualize the learned instance embeddings before bag-level aggregation using t-SNE (van der Maaten & Hinton, 2008). The comparison is conducted between ABMIL and ABMIL+BU-HIM on CAMELYON16. Both methods use the same test instances and identical t-SNE configurations to ensure a fair comparison. The tumor and normal labels are used only for visualization and are not involved in model training or representation learning.

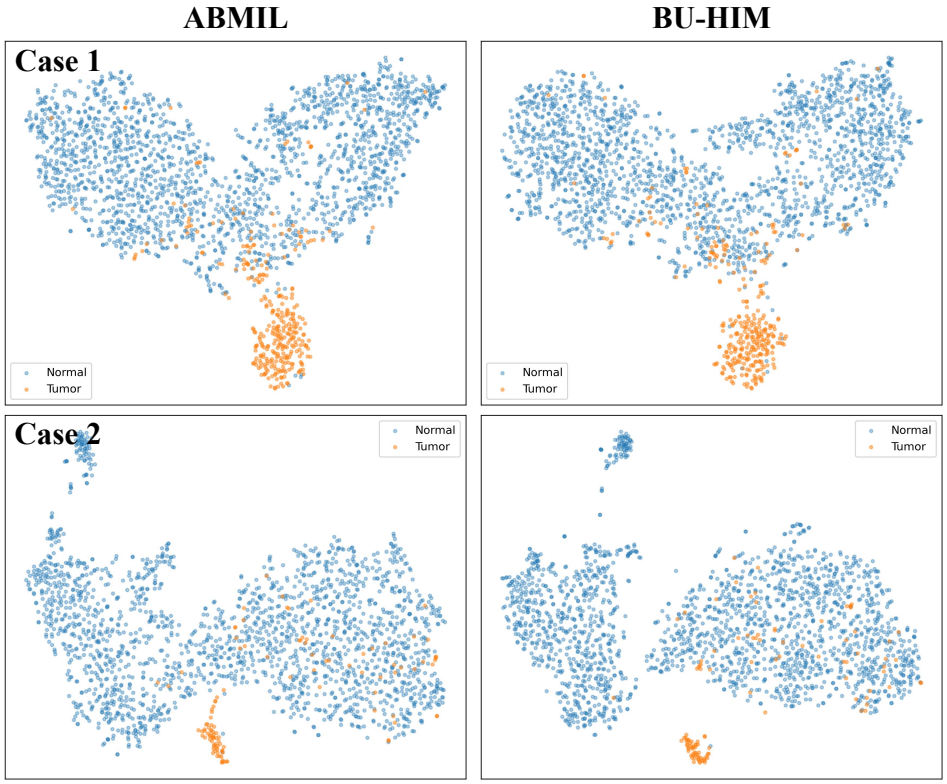

Figure 5: Comparison of instance feature distributions between ABMIL and BU-HIM in two representative CAMELYON16 cases.

As shown in Figure 5, the instance representations learned by ABMIL exhibit partial overlap between normal and tumor regions, especially near the class boundary. Some tumor instances are distributed within regions dominated by normal instances, indicating that the learned feature space still contains ambiguous representations.

After introducing BU-HIM, the projected instance representations show clearer class organization, with reduced inter-class overlap and more compact intra-class distributions in the displayed cases. This suggests that uncertainty-guided hard instance mining encourages the model to further optimize instances whose importance estimation remains uncertain, thereby improving the discriminative structure of instance representations.

Since t-SNE is a nonlinear dimensionality-reduction technique that can be affected by initialization and hyperparameter choices, these visualizations are considered qualitative observations rather than independent quantitative evidence. Together with the quantitative results in Table 1 and the attention localization analysis in Section 4.4, the feature visualization provides additional support that BU-HIM improves instance-level representation learning beyond deterministic attention-based aggregation.