WsiSSM: A Weakly Supervised Subset-Matching Framework for Unified Classification and Segmentation of Histopathology Whole Slide Images

Guangjian Zeng[1,2] **Siyuan Tao**[1,2] **Chengzhi Zhao**[3] **Xiaoqing Li**[4] **Na Tang**[5]
Wenting Huang[6] **Shenying Fang**[1,2]*

[1] School of Public Health and Emergency Management, Southern University of Science and Technology
[2] SUSTech Homeostatic Medicine Institute, Southern University of Science and Technology
[3] School of Basic Medical Sciences, Fudan University
[4] Department of Pathology, Nanfang Hospital
[5] Department of Pathology, Shenzhen People's Hospital
[6] National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College

Abstract

Multiple instance learning (MIL) has been increasingly used to analyze histopathology whole slide images (WSIs) and has achieved accurate diagnostic classification and a certain degree of automatic segmentation capability. However, segmentation performance without annotation remains suboptimal due to the immense size of WSIs, despite its essential role in assisting pathologists' diagnosis. Moreover, existing approaches treat classification and segmentation as independent tasks, with segmentation primarily serving as an interpretability section. To address this limitation, we propose a novel weakly supervised learning algorithm that unifies classification and segmentation within a framework. Specifically, we explore the interconnections between patches and introduce class activation maps in WSI analysis through the proposed Sparse-CAM backbone to activate the segmentation capability, and further enhance feature representation using the proposed subset-matching module to improve both classification and segmentation performance. The proposed method demonstrated superior performance in diagnosis and achieved state-of-the-art performance in segmentation compared to other MIL methods across multiple datasets. In addition, it offered high-precision segmentation that is sensitive to micro-tumor regions without pixel-level annotations. The code is publicly available at `https://github.com/GaryinDeep/WsiSSM`.

1 Introduction

As the diagnostic gold standard, histopathology relies on microscopic tissue analysis by experienced pathologists to assess suspected lesions. Hematoxylin and eosin (H&E) staining and immunohistochemistry (IHC) are primary techniques in histology used to highlight the fine structures of cells and tissues. Pathologists first assess the slide quality visually, then initially identify lesion areas under low-power microscopy. Upon locating a suspicious region, it is centered and examined in detail using high-power microscopy by the pathologist. Consequently, histopathological examination faces challenges such as strong subjectivity in diagnosis, high workload, and the potential to overlook

*Corresponding author: `fangsy@sustech.edu.cn`)

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

critical areas [21]. Advancements in digital scanning technologies have enabled the conversion of physical slides into high-resolution whole slide images (WSIs) that accurately capture the entire pathological tissue. This progress has significantly improved the efficiency of pathologists and driven the demand for intelligent WSI diagnostic tools [28, 31]. Numerous studies in recent years have demonstrated promising results by exploring deep learning applications in WSIs [35]. However, the large scale of WSIs (over 100,000 × 100,000 pixels at 40× magnification) presents a significant challenge for obtaining pixel-level annotations, leading most research to focus on weakly supervised learning that can be trained using only slide-level labels. Among weakly supervised methods for WSI classification, multiple instance learning (MIL) research is predominant. MIL reduces the dimensionality of WSIs by dividing them into smaller patches while preserving detailed tissue patterns, thus facilitating the development of complex deep learning architectures [23, 26, 41, 45]. Traditional MIL calculates a learnable attention score for each instance and performs weighted aggregation to obtain a bag-level representation at the WSI level called attention-based MIL. However, attention-based MIL focuses primarily on the impact of individual patches on the overall outcome, while ignoring their interconnections [24].

Recent advances in multiple instance learning (MIL) have explored graph-based models [17,24,29,30] and Transformer-based architectures [9, 10, 34, 43], which capture both the contribution of individual patches and their interrelationships via self-attention or neighborhood aggregation. These approaches have achieved near-pathologist-level diagnostic performance [1, 2]. However, applying Transformer-based MIL directly to whole-slide images (WSIs) remains prohibitively challenging. Gigapixel WSIs generate ultra-long patch sequences, where conventional self-attention becomes computationally infeasible. To mitigate this, sparse attention mechanisms have been introduced, substantially reducing complexity. Yet most variants, such as sliding-window [25] or block-sparse attention [12]), prioritize local context at the expense of global dependencies. In pathology, diagnostically relevant cues are often sparse and spatially scattered across a slide, meaning that the inability to capture both long-range context and fine-grained local dependencies, leading to attention shift phenomena where the model emphasizes different local patches to generate unstable representations and irrelevant or incomplete activated regions. This ultimately undermine the effectiveness of weakly supervised learning paradigms, as attention-based MIL fundamentally relies on attention maps to localize suspicious regions.

However, the issue of interpretability has long hindered the clinical translation of MIL, making it difficult to understand models' decision-making process [32]. Existing methods often rely on Grad-CAM [33] as a post-hoc visualization, which does not influence training and thus produces attention maps that are incomplete or clinically implausible. Directly integrating CAM into a sparse self-attention backbone can leverages the efficiency of sparse attention while simultaneously enforcing region-level activations during learning. This fusion not only guides the model to attend to diagnostically meaningful regions but also enhances classification itself by aligning optimization with clinically relevant localization. However, in practice, the risk of attention drift that the model generate unstable feature representations and false-positive regions may still persists during training in ultra-large whole-slide images.

For this purpose, we introduce a weakly supervised learning framework with the Sparse-CAM backbone and the **s**ub**s**et-**m**atching strategy for histopathology **WSI** classification and segmentation (WsiSSM). The proposed Sparse-CAM integrates CAM with a sparse self-attention, which improves the insufficient extraction of patches' interconnections and enhances the automatic segmentation capability of self-attention mechanisms. Furthermore, built upon a two-stage training scheme (detailed in Section 4.2), the subset-matching strategy reduces inconsistency of activated region by encouraging alignment of discriminative subsets across instances. This ensures that attention remains anchored to biologically relevant structures, reducing false-positive detections and improving both classification accuracy and weakly supervised segmentation quality. To demonstrate the effectiveness of our method, we conducted extensive evaluations across multiple datasets, including typing for breast cancer metastases and prostate cancer. Through comparisons with various SOTA MIL methods and ablation studies, we demonstrated the superior performance of WsiSSM in histopathology WSI classification and segmentation. This work provides a novel research perspective for MIL, establishes an evaluation standard of segmentation without pixel-level annotation, and enhances the clinical applicability of MIL.

The primary contributions of this paper can be summarized as follows: (1) We propose the Sparse-CAM backbone to improve the extraction of patches' interconnections and enhance the automatic

segmentation capability of self-attention mechanisms in WSI analysis. (2) We propose the subset-matching strategy to reduce the contrast between global CAM from the whole set and local CAMs from the unequal-length subsets, significantly improving the performance of both classification and segmentation without the need for pixel-level annotations. (3) Establishing new standards for MIL methods. We propose a quantitative assessment of suspicious region segmentation in MIL to convince pathologists and promote its application in clinical practice.

2 Related Works

2.1 MIL in Histopathology WSIs

MIL algorithms use only slide-level labels instead of detailed pixel-level annotations for training. MIL models can generally be divided into two groups: those where the final WSI predictions are directly derived from instance scores [4, 22, 42], and those where predictions are based on the aggregation of instance embeddings [11, 18, 23, 26]. The research trend has gradually shifted towards the embedding-based method due to its better performance [34, 40]. Many embedding-based MIL methods adopt the fundamental concept of AB-MIL [18], which involves calculating a learnable attention score for each instance and performing weighted aggregation to obtain a bag-level representation at the WSI level. DS-MIL [23] improved upon AB-MIL by calculating the attention score based on the cosine distance between features of an instance and the critical instance. Feng et al. [15] adopted a specific sampling strategy for sample selection and a transfer learning strategy for model training. CLAM [26] introduced a clustering-constraint loss to enhance the representation of the embedding feature. DTFD-MIL [45] introduced the concept of pseudo-bags and proposed a feature distillation strategy to fuse representations at the pseudo-bag level. UNI [8] improved the representation of the feature extractor using a pre-trained image encoder trained through self-supervised learning on publicly available histopathology datasets. While these methods have sought to enhance the analytical performance of WSIs, they still fall short in capturing the informative interactions between patches. Graph neural network (GNN) and Transformer excel at encoding representations and capturing feature interactions and have attracted widespread attention in MIL algorithms. With the application of Transformer, Trans-MIL [34] encoded the mutual interconnections between instances. GTP [49] fused a graph-based representation of a WSI and a vision Transformer (ViT) for processing pathology images. HIPT [7] introduced a new ViT architecture that leverages the natural hierarchical structure inherent in WSIs, using two levels of self-supervised learning to learn high-resolution image representations. Prov-GigaPath [43] adapted the newly developed LongNet [14] method to digital pathology, and achieved SOTA performance on various digital pathology tasks. For the GNN application, Zhao et al. [47] employed a graph convolutional network to learn the bag-level representation and perform the classification. HEAT [6] presented a new heterogeneous-graph edge attribute Transformer to leverage the edge and node heterogeneity during message aggregating. WiKG [24] proposed a novel dynamic graph representation algorithm that conceptualizes WSIs as a form of the knowledge graph structure. However, due to their mechanisms, GNN-based and Transformer-based MILs still have difficulty in performing automatic segmentation, which restricts their application in the real world.

2.2 Weakly Supervised Learning in Natural Images

Significant differences exist between the weakly supervised learning methods for natural images and histopathology WSIs. In natural images, weakly supervised learning primarily focuses on CAM generation and segmentation without pixel-level annotation. Zhou et al. [50] revisited the global average pooling (GAP) layer, highlighting how it explicitly enhances the localization ability of convolutional neural networks (CNNs) and proposed the concept of CAM. SEC [20] introduced a novel loss function for the weakly-supervised training of semantic image segmentation models based on three guiding principles: seeds, expansion, and constraints. Puzzle-CAM [19] addresses the main limitation of CAM, which only focuses on the most discriminative parts of the object, by minimizing the difference between features from individual patches and the whole image. CPAL [37] presented a context prototype-aware learning strategy to accurately capture the intra-class variations of object features through context-aware prototypes and promote adaptation to the semantic attributes of various instances. Despite the rapid development of weakly supervised learning in natural images, the extremely large size of histopathology WSIs makes it challenging to directly apply existing weakly supervised learning techniques, which were originally designed for natural or smaller medical images, to WSIs.

3 Methodology

3.1 WsiSSM Pipeline

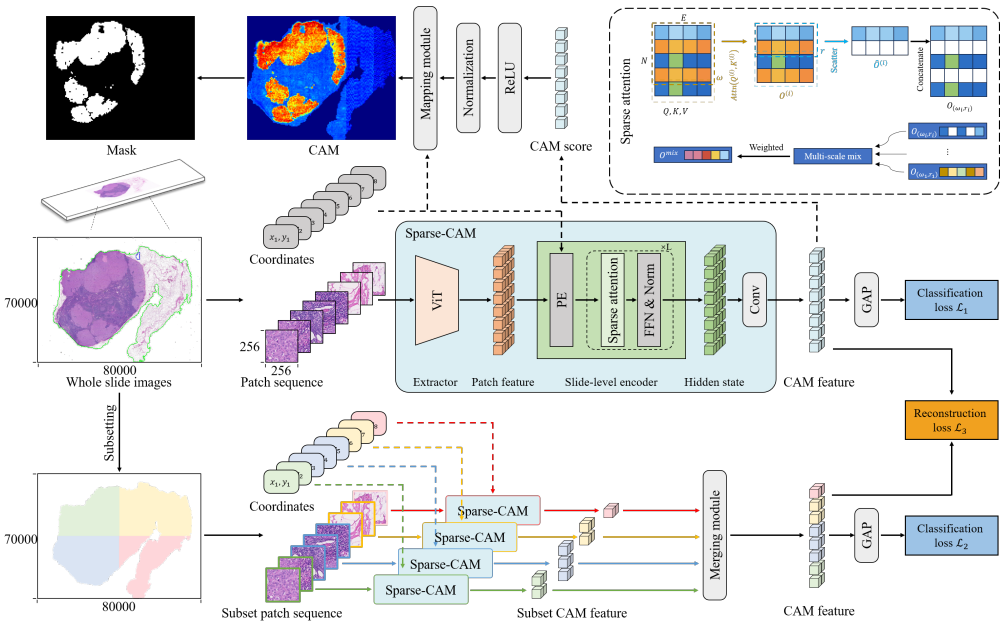

Figure 1: The pipeline of our proposed WsiSSM for WSI analysis, including rough segmentation, patch feature extraction, self-attention representation learning, and subset-matching module. CAM scores from WsiSSM are mapped to a CAM of the same size as the WSI according to the corresponding coordinates.

WSI processing. For WSI, our pipeline (fig. 1) begins with automated rough segmentation of the tissue regions. The WSI is converted from RGB to the HSV color space under 64× downscale resolution. A binary mask for the foreground tissue regions is computed by thresholding the saturation channel of the image after applying median blurring to smooth the edges, followed by additional morphological closing to fill small gaps and holes. The approximate contours of the detected foreground objects are then filtered based on an area threshold. After rough segmentation, we resized the WSI to a standard resolution of 0.5 μm per pixel (MPP) using the pyvips library and cropped non-overlapping patches $x \in \mathbb{R}^{256\times256}$ from within the segmented foreground contours while preserving their corresponding coordinates $p \in \mathbb{R}^2$ which represents the horizontal and vertical coordinates on the WSI. Following patching, we use the feature extractor of Prov-GigaPath, a foundation ViT model pretrained with standard DINOv2 [27] as an extractor to obtain a low-dimensional feature representation $g \in \mathbb{R}^M$ for each image patch of each WSI, where M is the dimension of the extractor's hidden state.

Sparse-CAM. The slide-level encoder is designed to capture the interconnections within the low-dimensional feature map of WSI $G \in \mathbb{R}^{N\times M}$. We project the feature map into an encoding space of dimension $E = 768$ through the encoding layer and perform 2D sine-cosine position embedding using patch coordinates $P \in \mathbb{R}^{N\times2}$, where N is the number of patches. The sine-cosine position embedding is computed as follows:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/(E/2)}) \tag{1}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/(E/2)}) \tag{2}$$

where pos is the position and i is the specific dimension $(2i \leq \frac{E}{2}, 2i+1 \leq \frac{E}{2})$. For (pos_1, pos_2) of the nth patch's coordinates, the 2D sine-cosine position embedding is computed as follows:

$$PE^{2D} = Concat(PE_{pos_1}, PE_{pos_2}) \tag{3}$$

Since conventional self-attention modules are not applicable for extremely long pathological patch sequences, we introduce a sparse attention to greatly reduce the computational complexity and calculate self-attention scores on patch sequences. The query–key–value matrices $(Q, K, V) \in \mathbb{R}^{N\times E}$ are partitioned into contiguous blocks of size ω:

$$(Q, K, V) \to \{(Q^{(i)}, K^{(i)}, V^{(i)})\}_{i=1}^{N/\omega} \tag{4}$$

The sparse self-attention is then computed as

$$O^{(i)} = softmax(\frac{Q^{(i)}K^{(i)T}}{\sqrt{E}})V^{(i)} \tag{5}$$

Within each block, we subsample tokens at interval γ, yielding

$$\tilde{O}^{(i)} = O^{(i)}[:: \gamma] \tag{6}$$

To maintain alignment with the original sequence length, outputs are scattered back according to the sampling stride:

$$\hat{O}^{(i)}[j] = \begin{cases} \tilde{O}^{(i)}[j], & j \; mod \; \gamma = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The outputs from all blocks are concatenated to form

$$O = Concat(\hat{O}^{(1)}, \hat{O}^{(2)}, ..., \hat{O}^{(N/\omega)}) \tag{8}$$

To simultaneously model local and global dependencies, multiple (ω, γ) pairs are employed, producing a family of sparse attentions $\{O_{(\omega_i,\gamma_i)}\}_{i=1}^{k}$. The final representation is a normalized weighted sum:

$$O^{mix} = \sum_{i=1}^{k} \frac{a_j}{\sum_{j=1}^{k} a_j} O_{(\omega_i,\gamma_i)}, a_j = \sum_{i=1}^{N} exp(W_a O_{(\omega_i,\gamma_i),i} + b) \tag{9}$$

where $W_a \in \mathbb{R}^{1\times E}$. In implementation, window sizes follow a geometric sequence $\omega = \{4, 8, \ldots, N\}$, with stride values $\gamma = \{1, \ldots, \omega/4\}$.

Given the 1D convolution layers θ with kernel size of K =1 as the classifier, we can generate CAM feature $A \in \mathbb{R}^{N\times C}$ that are the collection of CAM scores for all of the classes from hidden state of the slide-level encoder $F \in \mathbb{R}^{N\times E}$, where C is the number of classes:

$$A_i = \sum_{k=0}^{K-1} \theta_k F_{i+k} + b \tag{10}$$

where i denotes the convolution position. After aggregation through the global average pooling (GAP) module, we can obtain the logit $\hat{Y} \in \mathbb{R}^c$, which can be delineated as follows:

$$\hat{Y} = \sigma(GAP(A)) \tag{11}$$

The cross entropy loss is computed as follows:

$$\ell(\hat{Y}, Y) = -\sum_{c=1}^{C} Y_c log(\hat{Y}_c) \tag{12}$$

The logit $\hat{Y}$ participates in the calculations of classification loss $\mathcal{L}_1$:

$$\mathcal{L}_1 = \ell(\hat{Y}, Y) \tag{13}$$

Subset-matching module. From WSI of size $W \times H$, 4 non-overlapping regions of size $W/2 \times H/2$ are separated. Correspondingly, 4 unequal-length subsets $\{X^1, X^2, X^3, X^4\}$ are divided from patch sequence $X \in \mathbb{R}^{N\times 256\times 256}$ based on the patch coordinates' regions. Subsequently, each subset X^k is processed by the extractor, encoder, and classifier to obtain its subset CAM feature A^k. Finally,

all subset CAM features are merged into the new CAM feature A^s with the same patch order as the A according to the coordinates through the merging module. After aggregation through the GAP module, we can obtain the subset logit $\hat{Y}^s$ which participates in the calculations of subset classification loss $\mathcal{L}_2$:

$$\mathcal{L}_2 = \ell(\hat{Y}^s, Y) \tag{14}$$

To better enhance the local details of the model, we add reconstruction regularization, $\mathcal{L}_3$ to match A and A^s, $\mathcal{L}_3$ is computed as follows:

$$\mathcal{L}_3 = \|A - A^s\|_1 \tag{15}$$

Ultimately, the loss $\mathcal{L}$ of participation in propagation can be calculated as:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \alpha\mathcal{L}_3 \tag{16}$$

where α is the weight parameter. $\mathcal{L}_1$ and $\mathcal{L}_2$ ensure the stability of the classification, while $\mathcal{L}_3$ strengthen the automatic delineation of key regions and boundaries.

3.2 CAM of WsiSSM

Given the classifier θ, we apply the weights of the c-channel classifier as θ_c on feature map F to obtain the CAM score A_c of class c as follows:

$$A_c = \theta_c^T F \tag{17}$$

The generated CAM score is normalized by the maximum value $a'_k = (a_k - a_{min})/(a_{max} - a_{min})$. Subsequently, the mapping module combines the scores $A'_c \in \mathbb{R}^N$ and the corresponding coordinates $P \in \mathbb{R}^{N\times 2}$ to generates CAM I with the same size of $W \times H$ as the original WSI:

$$I(x,y) = \begin{cases} A'_{c,i}, & x \in [P_{i,1}, P_{i,1} + 256], y \in [P_{i,2}, P_{i,2} + 256] \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

Finally, the model prediction CAM is segmented into a binary mask with a threshold of 0.5.

4 Experiments

This section presents the performance of our proposed WsiSSM, comparing it with SOTA MIL methods across multiple datasets, and with a fully supervised method trained with pixel-level annotations to highlight WsiSSM's segmentation capability. Additionally, we conduct ablation experiments to explore the proposed methods and discuss the mechanism behind WsiSSM, including the subset-matching module.

4.1 Datasets

We evaluate our proposed WsiSSM on five multi-center histopathology WSI datasets, CAMELYON, PANDA, NCPDC and two external validation datasets.

CAMELYON consists of Camelyon16 [2] and Camelyon17 [1], which include lymph node tissue slides for detecting breast cancer metastases, with an average image size of approximately $10^5 \times 10^5$ pixels at $40\times$ magnification. CAMELYON16 consists of 270 slides for training and another 129 slides as a held-out official test set collected at the Radboud University Medical Center and the University Medical Center Utrecht in the Netherlands. Pixel-level tumor/normal annotations are available for all tissue areas in metastatic slides. On the other hand, CAMELYON17 comprises 1,000 slides collected from five medical centers across the Netherlands. Since slide-level labels for the 500 slides in CAMELYON17's official test set are not publicly available, only the training portion which consists of 500 slides with slide-level diagnoses was used, while only 50 slides have pixel-level region annotations. Combining CAMELYON16 and CAMELYON17 results in a dataset with 899 slides (593 negative, 306 positive). Before patching, we resized the slides in CAMELYON to a standard resolution of 0.5 MPP, i.e., 20× magnification, to ensure a unified scale.

We additional obtained two datasets on lymph node metastasis of breast cancer from two local hospitals: Chinese Academy of Medical Sciences and Peking Union Medical College (External A), and Nanfang Hospital (External B). At External A, 743 WSIs (73 tumor, 670 benign) were collected from 200 breast cancer patients in 2024. From this cohort, 19 slides from 11 randomly selected patients ($\sim 5\%$) were manually annotated by experienced pathologists. At External B, 564 WSIs (304 tumor, 260 benign) were collected from 61 patients in 2024. From this cohort, 12 slides from 7 randomly selected patients ($\sim 10\%$) were manually annotated by experienced pathologists.

PANDA [3] is the most extensive multi-center dataset on Gleason grading of prostate tissue, with a maximum resolution of 20× and an average image size of approximately $10^4 \times 10^4$ pixels. We selected benign and malignant slides, excluding those of low quality or insufficient resolution to support rough segmentation, resulting in a final set of 692 slides (360 negative, 332 positive). The specific IDs of slides included in this study will be stored in the code repository.

Given the scarcity of computational pathology datasets that provide both high-quality classification labels and pixel-level segmentation annotations, we will release NCPDC, a frozen-section dataset for primary breast cancer. In 2024, we collected 768 primary-tissue whole-slide images (WSIs) (334 tumor and 434 benign) from 670 patients. Slides were scanned in brightfield mode using the KFBIO KF-PRO-040-HI system, initially stored as KFB files at 20× magnification, and then converted to standard SVS format using KFbioConverter. From this cohort, six slides from six randomly selected patients in each test set were manually annotated by experienced pathologists using ASAP, with an average annotation time of approximately 4 hours per slide. All annotations were saved as XML files.

We used 10-fold Monte Carlo cross-validation [26], with each metric recorded as a percentage value accompanied by the margin of error for the 95% confidence interval.

4.2 Implementation Details

The proposed methodology was implemented using the Pytorch library in Python. We adopted a two-stage training strategy: first, the model was trained to convergence with $\mathcal{L}_1$, then further finetuned with $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ to enhance performance. For the hardware, three Nvidia RTX A6000 GPUs, each with 48 GB of graphics memory, were used. We extracted a 1536-dimensional feature vector for each patch through the extractor and froze the weights of the extractor during training. The batch size was set to 1, meaning each iteration processes only a single slide. We set the gradient accumulation with 32 iterations and the early stopping with patience of 5 to prevent overfitting. The number of epochs was set to 20. We used the AdamW optimizer with a learning rate of 2.5×10^{-4} and a weight decay of 0.05 for model weight. The learning rate of each parameter group was updated by the cosine annealing schedule during the training. The sensitivity and specificity were reported for all classification experiments due to their concern with the accuracy of a screening test relative to a reference standard. These metrics were determined by a threshold of 0.5. Additionally, the area under the curve (AUC) was reported for all classification experiments. In segmentation experiments, we considered the dice coefficient and intersection over union (IoU), two of the most commonly used semantic segmentation metrics. Additionally, accuracy was also reported for all segmentation experiments.

4.3 Comparison with Other Existing Works

Method	CAMELYON			PANDA			NCPDC			External A			External B		
	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
WOTC	93.80±6.20	98.70±1.30	96.60±3.20	–	–	–	–	–	–	–	–	–	–	–	–
WTC	62.80±9.10	98.50±6.00	81.00±5.90	–	–	–	–	–	–	–	–	–	–	–	–
CLAM (NatBiomedEng'2021)	78.59±4.78	97.10±1.88	93.44±1.84	95.41±2.95	82.20±4.81	97.46±1.34	96.84±1.96	98.11±3.07	99.46±0.42	72.73±3.48	82.59±2.71	82.31±1.17	91.11±1.69	95.88±1.29	96.23±0.46
TransMIL (NeurIPS'2021)	69.17±6.59	94.11±1.99	95.11±1.74	96.56±2.16	96.54±1.57	98.77±0.87	98.66±1.29	96.54±2.69	99.58±0.27	74.60±3.68	87.55±3.39	84.38±1.13	90.28±1.44	95.04±1.15	95.65±0.57
DSMIL (CVPR'2021)	55.60±6.29	94.75±2.22	82.16±1.97	92.23±3.33	87.51±3.34	95.91±1.51	93.49±3.56	98.02±1.62	98.28±1.07	76.35±2.88	89.08±1.46	86.26±1.24	92.72±0.81	96.17±0.54	96.54±0.35
DTFD-MIL (CVPR'2022)	82.83±1.72	95.80±1.83	88.87±2.03	97.59±1.49	87.66±4.09	98.52±0.88	97.15±1.63	99.49±1.16	99.63±0.20	70.58±3.03	90.75±1.24	82.35±1.27	94.22±1.39	93.96±1.07	96.69±0.38
Bayes-MIL (ICLR'2023)	93.21±2.38	96.46±1.27	93.26±2.35	**98.79±1.52**	98.44±0.64	99.39±0.76	97.88±1.46	98.60±1.16	99.84±0.19	87.67±1.20	97.79±1.18	92.87±0.63	97.71±0.48	98.04±0.50	97.98±0.19
MambaMIL (MICCAI'2024)	92.98±2.69	98.35±1.06	94.92±1.83	97.88±1.78	99.04±0.84	99.32±0.68	98.48±1.53	98.60±1.40	99.60±0.49	87.97±1.17	**98.90±2.80**	93.62±0.84	96.26±0.55	97.42±0.43	97.96±0.18
RRTMIL (CVPR'2024)	90.46±2.50	98.86±0.32	95.63±1.74	97.88±1.78	98.33±1.39	99.32±0.59	98.58±2.75	98.70±1.42	99.77±0.31	88.97±1.36	97.56±0.81	92.76±0.65	97.85±0.44	97.28±0.76	98.47±0.21
WiKG (CVPR'2024)	93.21±2.38	98.76±0.91	96.67±0.91	95.17±2.76	83.65±3.95	95.03±2.11	99.51±1.10	95.57±2.18	99.48±0.32	86.31±3.56	98.21±0.62	92.16±2.14	97.16±0.73	97.32±0.86	98.69±0.40
Prov-GigaPath (Nature'2024)	96.05±1.54	98.06±0.66	97.21±0.80	97.73±1.40	98.46±2.45	99.54±0.26	99.19±1.25	99.03±1.25	99.89±0.12	79.07±3.91	94.46±0.78	93.15±1.53	**98.77±0.51**	97.06±0.46	98.80±0.30
2DMamba (CVPR'2025)	92.99±2.05	98.60±0.46	94.93±1.93	97.98±1.74	98.33±1.39	99.32±0.59	98.18±2.09	98.14±1.72	99.61±0.47	88.25±1.99	96.26±0.71	89.22±1.30	96.80±0.61	97.39±0.78	98.39±0.25
M3amba (CVPR'2025)	94.38±2.48	92.75±2.99	97.00±0.65	97.88±2.30	99.12±0.63	99.34±0.67	98.79±2.09	99.30±1.12	98.93±1.18	87.66±1.42	89.71±1.96	93.02±0.53	96.61±1.27	96.62±0.92	98.22±0.15
ABMILX (NeurIPS'2025)	92.80±2.32	98.89±0.26	95.94±1.37	97.27±2.16	98.44±0.74	99.08±0.38	98.48±1.53	98.60±1.40	99.75±0.33	85.36±4.19	89.52±5.26	93.70±0.81	97.46±0.44	97.43±0.38	98.07±0.19
C2MIL (ICCV'2025)	91.55±2.42	97.29±1.89	95.17±1.58	96.97±0.63	99.02±0.63	98.76±1.11	95.15±3.71	98.60±1.61	99.32±0.75	84.05±2.27	88.67±1.24	89.99±0.95	97.42±0.73	98.01±0.66	98.37±0.31
BioX-CPath (CVPR'2025)	92.84±2.50	97.18±1.34	96.78±1.16	97.22±2.14	98.17±2.78	99.44±0.54	98.33±1.34	99.49±1.16	99.67±0.23	86.47±4.32	76.76±4.74	88.46±0.65	97.65±0.37	98.38±0.42	98.62±0.23
WsiSSM(ours)	**96.67±0.91**	**99.05±1.05**	**97.44±0.90**	97.44±1.94	**99.24±1.22**	**99.64±0.28**	**99.47±0.80**	**99.53±0.72**	**99.95±0.05**	**89.30±2.64**	96.82±0.66	**93.79±1.31**	97.13±0.83	**98.39±0.91**	**98.87±0.40**

Table 1: Classification performance [%] of various MIL methods across multiple datasets, The Symbol ± denotes the margin of error and the **BOLD** highlights the best performance.

We present the experimental classification results of the proposed methods, in comparison to a set of SOTA methods that include both foundational classics and recent top-conference publications.

These benchmark methods are: CLAM [26], TransMIL [34], DSMIL [23], DTFD-MIL [45], Bayes-MIL [13], MambaMIL [44], RRTMIL [39], WiKG [24], Prov-GigaPath [43], 2DMamba [46], M3amba [48], ABMILX [38], C2MIL [5], and BioXCPath [16]. The study also incorporated the pathologist benchmarks from [2]: the WOTC (without time constraint) and WTC (with a 180-minute time constraint) evaluations on the CAMELYON dataset. To ensure a fair comparison, all baseline methods utilize the same feature extractor and are evaluated from the experiments conducted using their official codes under the same settings. The classification results are presented in table 1. For the CAMELYON dataset, our method outperformed WOTC and achieved the best performance among all comparison methods with an AUC of 97.44%. For the PANDA dataset, our method is also in the leading position, achieving an AUC of 99.64%, demonstrating the strong representation utility of our architecture.

Method	CAMELYON			PANDA			NCPDC			External A			External B		
	Accuracy	Dice	IOU	Accuracy	Dice	IOU	Accuracy	Dice	IOU	Accuracy	Dice	IOU	Accuracy	Dice	IOU
CLAM (NatBiomedEng'2021)	91.76±1.06	20.85±3.10	11.70±1.90	91.37±0.73	31.78±0.98	18.90±0.70	84.81±1.60	19.92±4.78	11.20±2.94	67.26±1.99	10.77±2.53	5.73±1.42	92.41±0.98	44.06±3.44	28.37±2.82
TransMIL (NeurIPS'2021)	97.55±0.79	18.69±5.22	10.47±3.24	91.18±0.64	31.01±1.55	18.37±1.08	92.95±3.08	6.47±1.90	3.36±1.03	88.97±2.64	27.01±7.17	15.95±4.58	92.08±1.29	22.82±1.70	12.90±1.09
DSMIL (CVPR'2021)	96.02±1.14	6.91±0.90	3.58±0.48	90.61±0.80	12.23±0.28	6.51±0.16	93.43±3.16	16.06±8.18	9.13±5.08	88.35±2.23	8.68±1.35	4.55±0.73	92.08±1.43	10.75±1.11	5.69±0.62
DTFD-MIL (CVPR'2022)	95.89±1.17	2.56±0.52	1.30±2.66	90.33±0.84	6.39±0.41	3.30±0.22	93.53±2.99	15.56±3.67	8.51±2.26	87.88±2.34	1.86±0.29	0.94±0.15	91.50±1.44	4.83±0.69	2.48±0.36
Bayes-MIL (ICLR'2023)	96.89±0.90	39.52±4.05	24.77±3.17	91.46±0.82	25.64±2.67	14.75±1.79	88.00±2.69	42.33±2.26	26.89±1.83	90.14±2.00	27.54±2.88	16.02±1.92	92.33±1.36	21.70±2.10	12.20±1.32
MambaMIL (MICCAI'2024)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
RRTMIL (CVPR'2024)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
WiKG (CVPR'2024)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Prov-GigaPath (Nature'2024)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2DMamba (CVPR'2025)	96.60±1.05	27.29±3.94	15.90±2.62	90.89±0.80	16.60±1.39	9.06±0.83	86.09±3.27	24.85±3.61	14.27±2.39	89.20±2.16	21.21±2.22	11.89±1.40	92.80±1.25	28.58±2.03	16.70±1.37
M3amba (CVPR'2025)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
ABMILX (NeurIPS'2025)	96.38±1.04	26.04±4.08	15.08±2.71	91.39±0.78	24.03±1.32	13.66±0.85	87.61±2.88	39.93±3.18	25.03±2.51	87.50±2.05	29.50±2.52	17.34±1.67	93.04±1.21	39.94±2.82	25.02±2.19
C2MIL (ICCV'2025)	96.66±1.09	30.46±6.52	18.27±4.53	90.98±1.49	24.16±1.74	13.76±1.12	87.41±3.12	41.87±3.47	26.58±2.76	86.67±2.38	10.60±3.90	5.68±2.18	91.64±1.27	27.22±5.28	15.95±3.61
BioX-CPath (CVPR'2025)	97.13±1.11	62.63±1.40	48.23±1.48	90.41±0.67	20.43±2.13	11.41±1.34	69.69±2.12	48.44±8.11	32.59±6.68	92.02±1.50	54.75±4.65	38.53±7.99	96.55±0.87	75.58±1.90	61.80±1.51
WsiSSM(ours)	**98.73±0.41**	**81.98±2.15**	**69.57±3.11**	**96.05±0.52**	**79.26±2.18**	**65.74±3.01**	**91.86±1.66**	**75.22±4.20**	**60.59±5.11**	**94.51±0.54**	**75.50±4.68**	**61.06±6.23**	**97.14±0.47**	**80.37±1.23**	**67.21±1.71**

Table 2: Segmentation performance [%] of various MIL methods across multiple datasets. The Symbol ± denotes the margin of error and the **BOLD** highlights the best performance.

On the other hand, we present the experimental segmentation results of the proposed methods in comparison to existing MIL methods: (1) CLAM [26], normalizes the attention scores directly from the attention module and treats them as heatmap scores. (2) TransMIL [34], after averaging the multi-head self-attention matrices, the attention vector corresponding to the [CLS] token is extracted and normalized as the heat map scores. (3) DSMIL [23] and (4) DTFD-MIL [45], normalizes the attention scores as heatmap scores. (5) Bayes-MIL [13], follows the official setting by perform 16 attention samplings for the same WSI, averaging and normalizing the attention scores as heatmap scores. (6) 2DMamba [46], (7) ABMILX [38] and (8) C2MIL [5] , normalizes the attention scores as heatmap scores. (9) BioXCPath [16], traces the SAGPooling attention scores and aggregates them across layers as node importance scores, and then normalizes them as heatmap scores. For other methods, the official implementations did not support extracting attention scores, and their original papers did not present segmentation prediction heatmaps. After the same heatmap score mapping step as WsiSSM, the heatmaps of all the compared methods are segmented into a binary mask with a threshold of 0.5. The segmentation results are presented in table 2. Our method demonstrates breakthrough segmentation ability without pixel-level annotations and achieves SOTA segmentation performance compared to existing MIL methods. On the CAMELYON dataset, it reaches a dice coefficient of 81.98% and an IoU of 69.57%. On the PANDA dataset, it achieves a dice coefficient of 79.26% and an IoU of 65.74%. Due to the large background region in WSIs, the accuracy across different methods shows minimal variation. fig. 4 and fig. 3 (Supplementary) presents the segmentation predictions of WsiSSM and other MIL methods on the test sets of the CAMELYON and PANDA datasets. WsiSSM correctly locates the suspicious region in the WSIs, and even for micro regions, our method can still delineate the corresponding area, while other MIL methods fall short.

4.4 Ablation Study

Method	CAMELYON			PANDA		
	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
$\mathcal{L}_1$	96.08±2.38	98.27±0.84	96.57±1.50	96.67±1.97	97.07±1.77	99.18±0.60
$\mathcal{L}_1$+$\mathcal{L}_2$	96.20±2.86	98.68±1.13	96.83±1.71	95.22±2.67	98.94±1.86	99.23±0.60
$\mathcal{L}_1$+$\mathcal{L}_2$+$\mathcal{L}_3$	**96.67±0.91**	**99.05±1.05**	**97.44±0.90**	**97.44±1.94**	**99.24±1.22**	**99.64±0.28**

Table 3: The classification results [%] of the ablation study in training regularizations on the CAMELYON and PANDA datasets. The Symbol ± denotes the margin of error and the **BOLD** highlights the best performance.

To verify the effectiveness of the proposed module, ablation experiments were conducted on the CAMELYON and PANDA datasets. We evaluated three training regularizations on performance. As shown in table 3, regularization $\mathcal{L}_3$ yields the most significant gains, with a +0.61% improvement in AUC on the CAMELYON dataset and +0.41% on the PANDA dataset.

For segmentation performance (table 4), regularization $\mathcal{L}_3$ also yields the most significant boost, with a +42.42% improvement of dice and +44.41% improvement of IOU on the CAMELYON dataset, and a +7.91% improvement of dice and +10% improvement of IOU on the PANDA dataset. Specific examples of segmentation improvements are shown in fig. 6 (Supplementary).

Method	**CAMELYON**			**PANDA**		
	Accuracy	**Dice**	**IOU**	**Accuracy**	**Dice**	**IOU**
$\mathcal{L}_1$	85.64±1.02	34.03±6.97	20.90±5.20	94.81±0.71	65.86±4.59	49.41±5.18
$\mathcal{L}_1$+$\mathcal{L}_2$	88.32±1.20	38.89±7.49	24.63±6.01	95.40±0.69	71.35±4.10	55.74±5.03
$\mathcal{L}_1$+$\mathcal{L}_2$+$\mathcal{L}_3$	**98.73±0.41**	**81.98±2.15**	**69.57±3.11**	**96.05±0.52**	**79.26±2.18**	**65.74±3.01**

Table 4: The segmentation results [%] of the ablation study in training regularizations on the CAMELYON and PANDA datasets. The Symbol ± denotes the margin of error and the **BOLD** highlights the best performance.

To assess the impact of subset quantity, we evaluated k-subsets' (k=4, 9, 16) performance (table 5 and fig. 7 in Supplementary). While subset-matching consistently improves performance, the 4-subset configuration performs best, indicating that excessive subdivision may lead to over-reliance on local information.

K-subset	**Classification**			**Segmentation**		
	Sensitivity	**Specificity**	**AUC**	**Accuracy**	**Dice**	**IOU**
4	**96.67±0.91**	**99.05±1.05**	**97.44±0.90**	**98.73±0.41**	**81.98±2.15**	**69.57±3.11**
9	94.87±1.17	98.39±1.62	97.01±1.12	97.38±0.87	56.49±4.92	39.66±4.80
16	94.33±1.48	98.79±1.62	97.18±0.99	92.79±0.57	46.75±7.13	31.01±6.15

Table 5: The classification and segmentation results [%] of the ablation study in number of subsets on the CAMELYON dataset. The Symbol ± denotes the margin of error and the **BOLD** highlights the best performance.

To verify the effectiveness of Sparse-CAM, we replaced various backbone components (table 6). MLP underperforms due to its inability to capture interconnections of patches, while Sparse-CAM achieves the best performance, highlighting the superiority of our proposed backbone.

Backbone	**Classification**			**Segmentation**		
	Sensitivity	**Specificity**	**AUC**	**Accuracy**	**Dice**	**IOU**
MLP + Average pooling	28.61±5.40	96.94±2.35	57.98±3.73	–	–	–
MLP + CAM	38.85±4.29	90.77±3.89	60.46±3.16	**96.63±1.09**	27.37±5.89	16.09±3.39
Block Sparse Attention [12]+ CAM	91.42±3.07	98.95±0.75	93.38±2.26	81.71±1.06	29.16±6.48	17.37±4.58
Sparse-CAM (ours)	**96.08±2.38**	**98.27±0.84**	**96.57±1.50**	85.64±1.02	**34.03±6.97**	**20.90±5.20**

Table 6: The classification and segmentation results [%] of the ablation study in backbone on the CAMELYON dataset. The Symbol ± denotes the margin of error and the **BOLD** highlights the best performance.

To clarify the impact of threshold selection on segmentation performance, we calculated and compared the evaluation metrics across different thresholds (fig. 2).

To explore the mechanism behind the subset-matching strategy, based on the two-stage training strategy, where the model first converges with regularization $\mathcal{L}_1$and is then enhanced with regularization $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$, we extracted heat maps of the whole set and four subsets from models trained with only $\mathcal{L}_1$ and with $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ (fig. 3). Without the effect of reconstruction loss $\mathcal{L}_3$, the heat maps of four subsets show different contrasts in subset regions (upper left), while under the regularization of $\mathcal{L}_3$, the heat maps of four subsets exhibit consistent contrast across subset regions (lower left), which helps to improve the overall segmentation performance of the whole set.

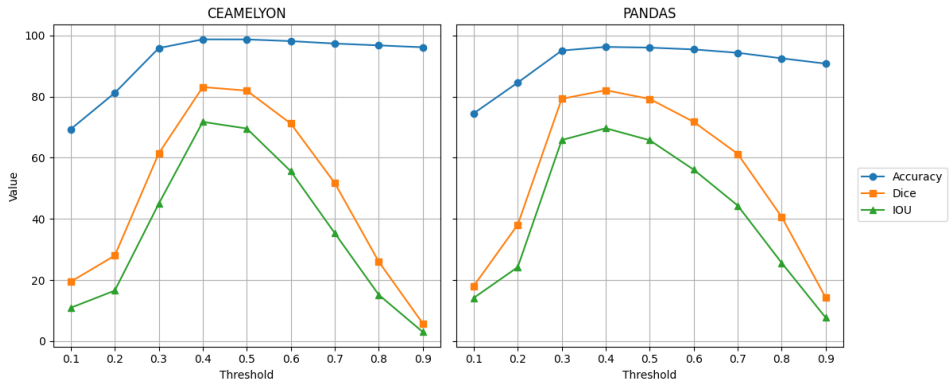

Figure 2: Impact of threshold on segmentation metrics.

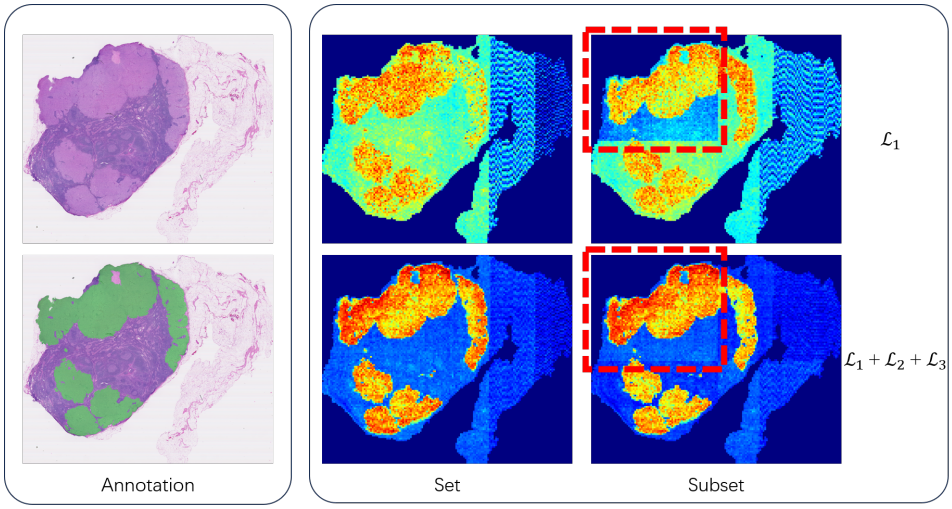

Figure 3: Visualization of WsiSSM's mechanism. Annotation is shown on the left, prediction with the whole subset in the middle, and prediction with four subsets on the right. The first row of the heatmaps presents the prediction of the model trained with only regularization $\mathcal{L}_1$, and the second row presents the prediction of the model trained with regularization $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$.

5 Conclusion

In this paper, we propose WsiSSM, a weakly supervised learning framework designed with Sparse-CAM backbone and subset-matching strategy for the classification and segmentation of histopathological WSIs. This approach aims to improve performance of both classification and segmentation tasks. Through extensive experiments across multiple datasets, we demonstrated the effectiveness and superior performance of WsiSSM. Consequently, WsiSSM exhibits considerable potential and is expected to gain the confidence of pathologists, paving the way for broader clinical application.

Acknowledgements

This work was supported by the Shenzhen Medical Academy of Research and Translation (Grant No. B2603003), the leading Talents Program of Guangdong Province (Grant No. 2023CX10Y002), and the National Health Commission Key Laboratory of Nuclear Technology Medical Transformation (Grant No. 2025HYX026).

References

[1] Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A.: From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE transactions on medical imaging **38**(2), 550–560 (2018)

[2] Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)

[3] Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., Van Boven, H., Vink, R.: Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature medicine **28**(1), 154–163

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

[5] Cen, M., Zhuang, Z., Zhang, Y., Zeng, M., Magnier, B., Yu, L., Zhang, H., Wang, L.: C2mil: Synchronizing semantic and topological causalities in multiple instance learning for robust and interpretable survival analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24392–24401 (2025)

[6] Chan, T.H., Cendra, F.J., Ma, L., Yin, G., Yu, L.: Histopathology whole slide image analysis with heterogeneous graph representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15661–15670 (2023)

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

[9] Chen, R.J., Krishnan, R.G.: Self-supervised vision transformers learn visual concepts in histopathology. arXiv preprint arXiv:2203.00585 (2022)

[10] Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4025 (2021)

[11] Chen, Y., Shao, Z., Bian, H., Fang, Z., Wang, Y., Cai, Y., Wang, H., Liu, G., Li, X., Zhang, Y.: dmil-transformer: Multiple instance learning via integrating morphological and spatial information for lymph node metastasis classification. IEEE Journal of Biomedical and Health Informatics **27**(9), 4433–4443 (2023)

[12] Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509 (2019)

[13] Cui, Y., Liu, Z., Liu, X., Liu, X., Wang, C., Kuo, T.W., Xue, C.J., Chan, A.B.: Bayes-mil: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In: 11th International Conference on Learning Representations (ICLR 2023) (2023)

[14] Ding, J., Ma, S., Dong, L., Zhang, X., Huang, S., Wang, W., Zheng, N., Wei, F.: Longnet: Scaling transformers to 1,000,000,000 tokens. arXiv preprint arXiv:2307.02486 (2023)

[15] Feng, R., Liu, X., Chen, J., Chen, D.Z., Gao, H., Wu, J.: 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)

[16] Gallagher-Syed, A., Senior, H., Alwazzan, O., Pontarini, E., Bombardieri, M., Pitzalis, C., Lewis, M.J., Barnes, M.R., Rossi, L., Slabaugh, G.: Biox-cpath: Biologically-driven explainable diagnostics for multistain ihc computational pathology. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10372–10383 (2025)

[17] Hou, W., Huang, H., Peng, Q., Yu, R., Yu, L., Wang, L.: Spatial-hierarchical graph neural network with dynamic structure learning for histological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 181–191. Springer (2022)

[18] Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)

[19] Jo, S., Yu, I.J.: Puzzle-cam: Improved localization via matching partial and full features. In: 2021 IEEE international conference on image processing (ICIP). pp. 639–643. IEEE (2021)

[20] Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 695–711. Springer (2016)

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

[22] Lerousseau, M., Vakalopoulou, M., Classe, M., Adam, J., Battistella, E., Carré, A., Estienne, T., Henry, T., Deutsch, E., Paragios, N.: Weakly supervised multiple instance learning histopathological tumor segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23. pp. 470–479. Springer (2020)

[23] 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)

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

[25] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)

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

[27] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

[28] Pallua, J., Brunner, A., Zelger, B., Schirmer, M., Haybaeck, J.: The future of pathology is digital. Pathology-Research and Practice **216**(9), 153040 (2020)

[29] Pati, P., Jaume, G., Ayadi, Z., Thandiackal, K., Bozorgtabar, B., Gabrani, M., Goksel, O.: Weakly supervised joint whole-slide segmentation and classification in prostate cancer. Medical Image Analysis **89**, 102915 (2023)

[30] Pati, P., Jaume, G., Foncubierta-Rodriguez, A., Feroce, F., Anniciello, A.M., Scognamiglio, G., Brancati, N., Fiche, M., Dubruc, E., Riccio, D.: Hierarchical graph representations in digital pathology. Medical image analysis **75**, 102264 (2022)

[31] Pell, R., Oien, K., Robinson, M., Pitman, H., Rajpoot, N., Rittscher, J., Snead, D., Verrill, C., group, U.N.C.R.I.C.P.q.a.w., Driskell, O.J.: The use of digital pathology and image analysis in clinical trials. The Journal of Pathology: Clinical Research **5**(2), 81–90 (2019)

[32] Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE **109**(3), 247–278 (2021)

[33] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

[34] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)

[35] Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Medical image analysis **67**, 101813 (2021)

[36] Takahama, S., Kurose, Y., Mukuta, Y., Abe, H., Fukayama, M., Yoshizawa, A., Kitagawa, M., Harada, T.: Multi-stage pathological image classification using semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10702–10711 (2019)

[37] Tang, F., Xu, Z., Qu, Z., Feng, W., Jiang, X., Ge, Z.: Hunting attributes: Context prototype-aware learning for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3324–3334 (2024)

[38] Tang, W., Qin, R., Fang, H., Zhou, F., Chen, H., Li, X., Cheng, M.M.: Revisiting end-to-end learning with slide-level supervision in computational pathology. arXiv preprint arXiv:2506.02408 (2025)

[39] Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11343–11352 (2024)

[40] Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern recognition **74**, 15–24 (2018)

[41] Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y.: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature pp. 1–9 (2024)

[42] Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., Xu, W.: Camel: A weakly supervised learning framework for histopathology image segmentation. In: Proceedings of the IEEE/CVF International Conference on computer vision. pp. 10682–10691 (2019)

[43] Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y.: A whole-slide foundation model for digital pathology from real-world data. Nature pp. 1–8 (2024)

[44] Yang, S., Wang, Y., Chen, H.: Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In: International conference on medical image computing and computer-assisted intervention. pp. 296–306. Springer (2024)

[45] 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)

[46] Zhang, J., Nguyen, A.T., Han, X., Trinh, V.Q.H., Qin, H., Samaras, D., Hosseini, M.S.: 2dmamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3583–3592 (2025)

[47] Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X.: Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4837–4846 (2020)

[48] Zheng, T., Jiang, K., Xiao, Y., Zhao, S., Yao, H.: M3amba: Memory mamba is all you need for whole slide image classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15601–15610 (2025)

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

[50] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)

A Additional Visualizations

fig. 4 and fig. 5 presents the mask and heatmap predictions of WsiSSM and other MIL methods on the test sets of various datasets.

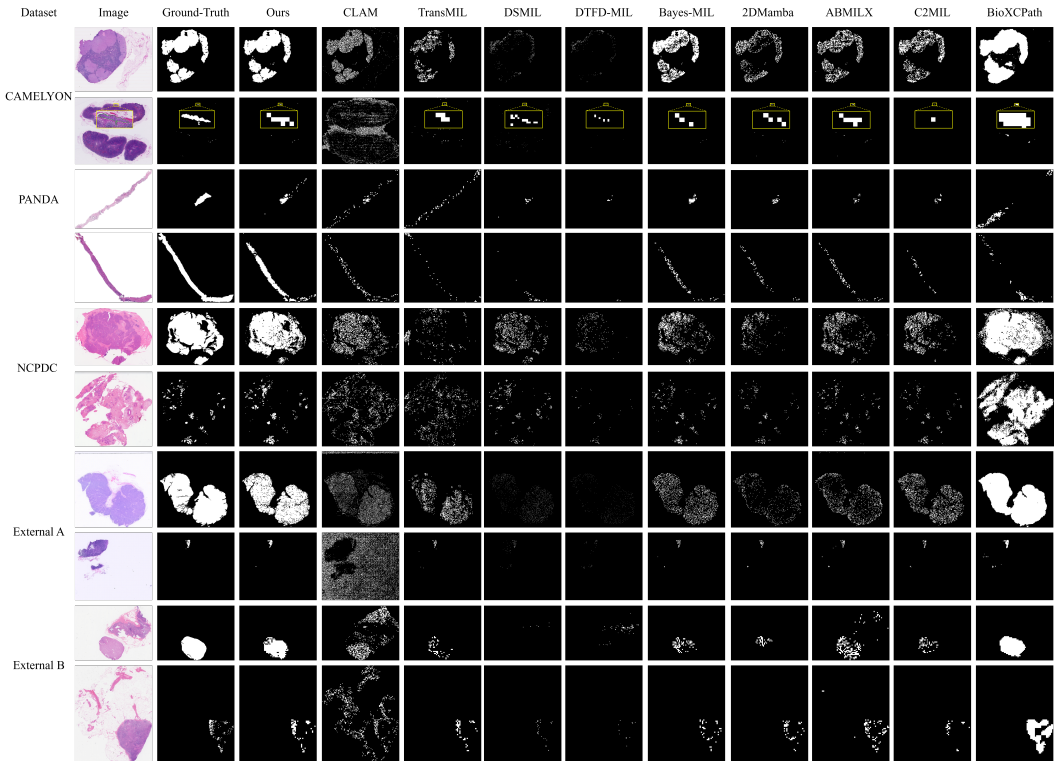

Figure 4: Visualization of segmentation mask from our and other MIL methods across multiple datasets. The cancer tissues are shown in white, and normal tissues are shown in black. To comprehensively demonstrate the segmentation performance, one large-lesion and one small-lesion case are selected from each dataset.

fig. 6 presents theheatmap and segmentation results in three training regularizations on CAMELYON and PANDA datasets .

fig. 7 presents the specific examples of segmentation predictions from model trained with k-subset in ablation study on CAMELYON datasets.

To further demonstrate the segmentation capabilities of our proposed weakly supervised (WS) approach, WsiSSM, which is trained without pixel-level annotations, we compared its performance against a fully supervised (FS) approach that uses pixel-level annotations, as proposed by Takahama et al [36]. The comparison results are presented in table 7. While WsiSSM does not perform as well as the fully supervised (FS) method on the CAMELYON dataset, it achieves better results on the PANDA dataset. Additionally, fig. 8 demonstrates that WsiSSM is more sensitive to micro-tumor regions and can effectively exclude benign tissues from the segmentation mask, thereby improving overall accuracy and reliability

Method	CAMELYON			PANDA		
	Accuracy	Dice	IOU	Accuracy	Dice	IOU
Takahama et al. (CVPR'2019)	**99.37±0.15**	**91.59±2.15**	**84.54±2.56**	92.40±0.86	45.50±5.72	29.76±4.84
WsiSSM(ours)	98.73±0.41	81.98±2.15	69.57±3.11	**96.05±0.52**	**79.26±2.18**	**65.74±3.01**

Table 7: Segmentation comparison [%] between our proposed weakly supervised method and a fully supervised approach on CAMELYON and PANDA datasets, The Symbol ± denotes the margin of error and the **BOLD** highlights the best performance.

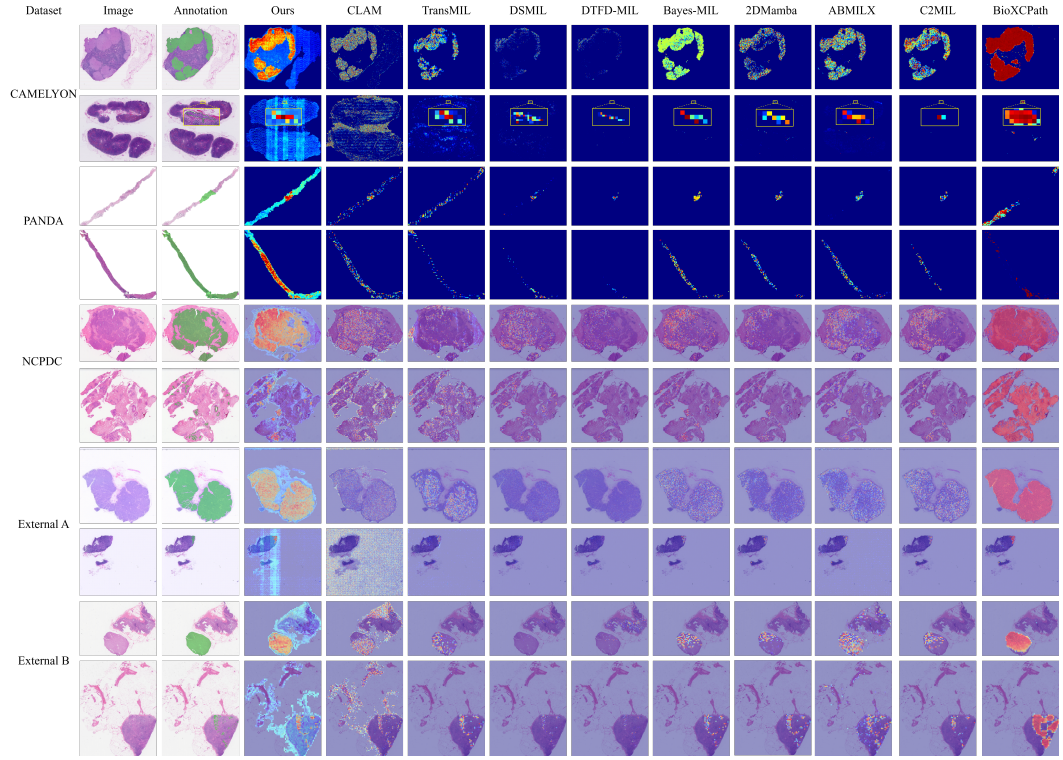

Figure 5: Visualization of heatmap from our method and other MIL methods across multiple datasets. Annotation's cancer tissues are shown in green, heatmap's cancer tissues are shown in red, and heatmap's normal tissues are shown in blue.

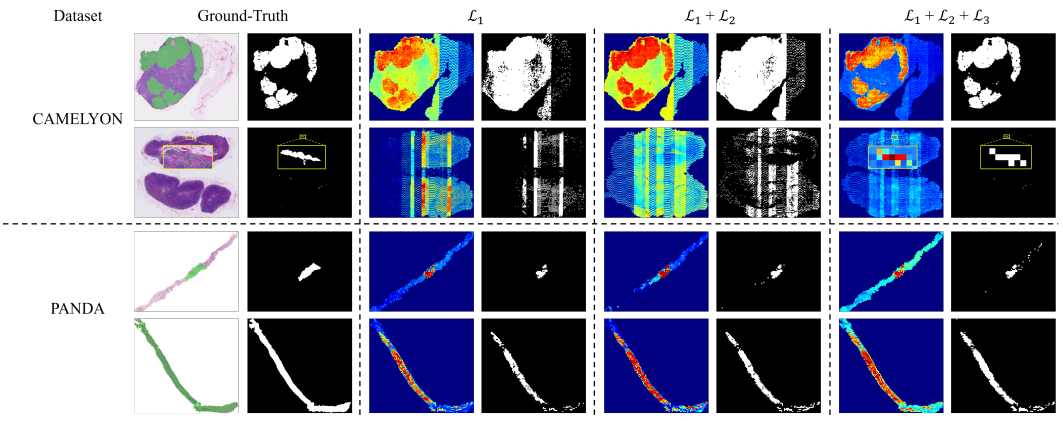

Figure 6: Visualization of heatmap and segmentation results in three training regularizations on CAMELYON and PANDA datasets. From left to right, the dashed lines separate the original WSI image, the ground truth, and the prediction of models trained with different settings. The predicted results include heatmaps and masks segmented with a threshold of 0.5.

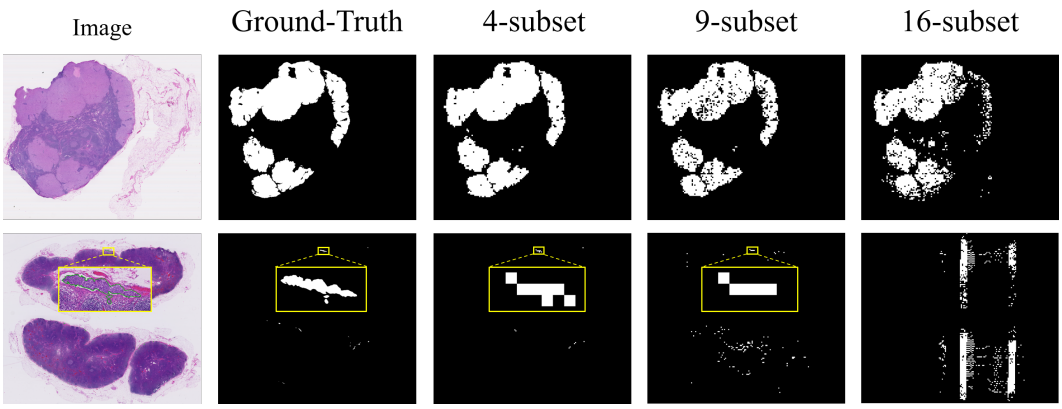

Figure 7: Visualization of segmentation mask from model trained with k-subset on CAMELYON datasets.The cancer tissues are shown in white, and normal tissues are shown in black.

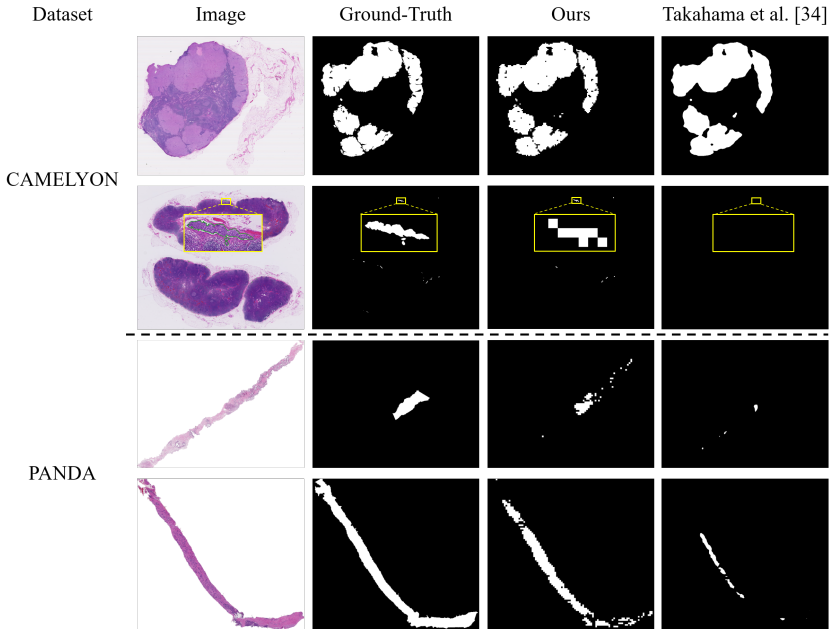

Figure 8: Visualization of segmentation mask from our methods and a fully supervised approach on CAMELYON and PANDA datasets. The cancer tissues are shown in white, and normal tissues are shown in black.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contribution of this article is precisely described in the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a limitation analysis in result section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Our paper does not include theoretical results

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a description of the data processing and model architecture in the paper, , and our code will be publicly available on GitHub.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use publicly available data and our code will be publicly available on GitHub upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: These are both detailed in the Sec. 4.2 and able to be referenced in the released code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We calculated confidence intervals on all results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include compute resources under training details in the Sec. 4.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We adhere to these guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss these implications in the introduction and conclusion.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data nor use scraped data.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the Data and packages are cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release our code along with its documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper uses publicly available datasets rather than directly interacting with any human subjects. Thus, this paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.