GroupMIL: Semantic Group Based Multiple Instance Learning for Whole Slide Image Analysing

Zhao Yao[1,2], **Zhenmi Xie**[1,2], **Mengxin Tian**[3], **Guoqing Wu**[4], **Yaonan Wang**[1,2] and **Min Liu**[1,2] *
[1]School of Artificial Intelligence and Robotics, Hunan University
[2]National Engineering Research Center of Robot Visual Perception and Control Technology
[3]Zhongshan Hospital, Fudan University
[4]College of Biomedical Engineering, Fudan University
{yaozhao24, septem, yaonan, liu_min}@hnu.edu.cn, tianmengxingyu@163.com,
guoqingwu@fudan.edu.cn

Abstract

Whole Slide Image (WSI) analysis faces challenges due to gigapixel resolutions and slide-level supervision. Multiple Instance Learning (MIL) serves as a pivotal method for this task. However, existing MIL frameworks often fail to exploit the inherent redundancy of tissue patterns or the semantic coherence among similar patches within a WSI. We propose *GroupMIL*, a novel framework that introduces a differentiable grouping mechanism into the MIL framework. This approach enables the automatic emergence of semantic segments using only slide-level labels. We specifically introduce a multi-stage grouping block and a hierarchical aggregator, which progressively fuse features within and across groups to construct a robust slide-level representation. Extensive experiments on multiple public datasets across cancer subtyping and survival prediction tasks demonstrate that *GroupMIL* consistently surpasses state-of-the-art performance.

1 Introduction

The digitalization of pathology through Whole Slide Images(WSIs) has revolutionized microscopic tissue analysis by applying deep learning technologies[Niazi *et al.*, 2019]. Recently, computer-Aided Diagnosis(CAD) has achieved significant success across diverse clinical applications such as cancer screening[Sakaguchi *et al.*, 2025], tumor grading[Campanella *et al.*, 2019], prognosis analysis[Tian *et al.*, 2024], gene mutant prediction[Ning *et al.*, 2025]. Despite these achievements, the gigapixel resolution of WSIs poses a primary challenge, as it renders conventional end-to-end analysis methods infeasible[Pati *et al.*, 2023].

To overcome this, Multiple Instance Learning (MIL) has emerged as the standard solution to WSI analysis[Wang *et al.*, 2016]. The typical MIL pipeline involves extracting patches from the WSI, generating embeddings via a pretrained feature extractor, and utilizing an aggregation module to aggregate these embeddings into a slide-level representation for tasks such as cancer subtyping and survival prediction[Kanavati *et al.*, 2020]. The evolution of WSI analysis is largely defined by the strategy used to model the relationship between patches. Initial research formalized MIL by learning a Bernoulli distribution over patch labels within a WSI using neural networks[Ilse *et al.*, 2018], as shown in Fig.1(a). Subsequent advancements, such as DSMIL[Li *et al.*, 2020], introduced non-local attention mechanisms to compute the similarity between the most discriminative patch and all other instances. To address overfitting, ACMIL[Zhang *et al.*, 2023b] implemented a strategy of randomly masking high-attention patches to redistribute focus to the remaining instances. These models assign the attention score to each patch in isolation and determine its contribution to the final diagnosis. While computationally efficient, these methods

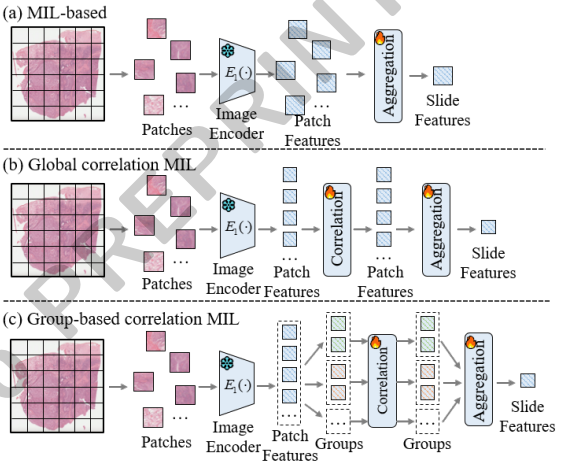

Figure 1: Comparison of our *GroupMIL* with existing MIL- and global correlation MIL-based methods. (a) MIL-based methods treat instances as independent and identically distributed, calculating scores for individual patches without considering correlation context. (b) Global correlation MIL generate the correlation among all patches and lacks explicit modeling of tissue subgroups. (c) Group-based correlation MIL (Ours) adaptively clusters patches into semantically coherent groups. It first calculate local correlations within each group and then aggregate all groups to form a robust and efficient slide-level representation.

*Corresponding author

represents a significant oversimplification of clinical pathology. Specifically, it ignores the morphological context and the semantic dependencies inherent in tissue structures. As an improvement, recent research has shifted toward modeling global correlations among patches, as shown in Fig.1(b). Frameworks such as TransMIL[Shao *et al.*, 2021] and its variants[Zhang *et al.*, 2023a] utilize Transformer-based architectures to establish a global patch-interaction model. By leveraging self-attention, these methods allow each patch to interact with all other patch in the WSI, effectively capturing long-range dependencies across the slide. However, global correlation models treat each patch as a distinct entity, thereby overlooking the inherent semantic redundancy of tissue where thousands of patches often represent identical biological components (e.g., a tumor region, a stroma region).

Inspired by [Xu *et al.*, 2022], we argue that the correlation between patches should be modeled through coherent groups derived from intrinsic feature similarity. In a WSI, patches belonging to the same tissue phenotype exhibit high feature redundancy. Therefore, an effective MIL framework should automatically aggregate these similar instances into distinct semantic units. By transitioning from global to group-based correlation, the model may explicitly captures the tissue's structural organization, yielding a more robust and semantically meaningful slide-level representation.

Based on this motivation, we proposed *GroupMIL*, a MIL framework designed to organize the unstructured "bag" of patches into semantic groups.The main difference between *GroupMIL* and existing paradigms are summarized in Fig.1(c). The core innovation lies in the introduction of learnable Semantic Group Tokens coupled with a dynamic assignment mechanism. Specifically, We introduce Adaptive Multi-stage Grouping module to transform the raw sequence of patch tokens into a structured hierarchy of Semantic Group Tokens. This process operates across multiple stages, progressively transitioning from fine-grained to coarse-grained semantic representations. Further, rather than relying solely on the final layer's output, *GroupMIL* employs Hierarchical Semantic Aggregation to preserve multi-scale context. By leveraging an attention-based fusion mechanism, HSA dynamically integrates the representative tokens from each stage, ensuring that the final slide-level embedding captures comprehensive information across multiple levels of granularity.

The main contributions of this work are summarized as follows:

(1) Adaptive Multi-stage Grouping: By adaptively clustering patches into semantic group tokens based on feature similarity, our model explicitly captures the latent histological structures of WSIs, transforming thousands of unorganized patches into a structured set of tissue phenotypes.

(2) Hierarchical Semantic Aggregation: We propose a feature integration strategy that captures the multi-scale information of pathological data. By combining the Adaptive Multi-stage Grouping with Hierarchical Semantic Aggregation, *GroupMIL* effectively preserves and weights information across different levels of granularity. This allows the model to leverage both high-resolution details and coarse-grained architecture to form a comprehensive WSI-level representation.

(3) Superior Performance across Extensive Clinical Tasks: We demonstrate the effectiveness and generalization of *GroupMIL* through extensive experiments on multiple datasets. Our model consistently outperforms state-of-the-art MIL frameworks across diverse tasks, including cancer subtyping and survival prediction, confirming its robustness and effectiveness in computational pathology analysing.

2 Method

In this section, we first provide a brief review of MIL and then introduce the overall *GroupMIL* architecture. Subsequently, we provide a detailed elucidation of its two core components: Adaptive Multi-stage Grouping and Hierarchical Semantic Aggregation.

2.1 Preliminaries

MIL servers as the predominant weakly supervised learning method employed to address the computational challenges posed by the gigapixel size of WSIs. Within the MIL paradigm, a WSI is considered as a bag of instances, $X = x_1, x_2, ..., x_K$. The bag-level label Y can be expressed as:

$$Y = \begin{cases} 1, & \text{if } \sum_i y_i > 1 \\ 0, & otherwise \end{cases} \tag{1}$$

where y_i is the latent label of instance x_i. To obtain a holistic representation of a WSI, two predominant methodologies are commonly employed. The first category assigns pseudo-labels to each instance and subsequently aggregates information from the top-k instances. The second, feature-embedding-based approach, extracts features for each instance using a pre-trained model and then integrates them, typically via an attention mechanism. As illustrated in Fig.1, these methods can be categorized into two main approaches. The first operates under the Independent and Identically Distributed (i.i.d.) assumption, utilizing attention mechanisms to aggregate features from all patches in isolation. The second approach leverages global correlation, performing exhaustive feature interactions across all patches to generate WSI representations. In contrast, our proposed *GroupMIL* introduces a hierarchical strategy: it first partitions patches into semantically coherent groups and subsequently calculates intra-group correlations to refine the local feature representations before final aggregation.

2.2 Overview of GroupMIL

The proposed *GroupMIL* framework, shown in Fig.2, is designed to perform hierarchical progressive grouping of visual concepts of WSI. Similar to many model paradigms[Dosovitskiy *et al.*, 2020][He *et al.*, 2015], *GroupMIL* follows a hierarchical multi stage design. At each stage, we first initialize learnable group tokens and generate global information with the WSI patch tokens using self-attention. Subsequently, the learned group tokens are employed to cluster the patch tokens, aggregating similar patch tokens together via

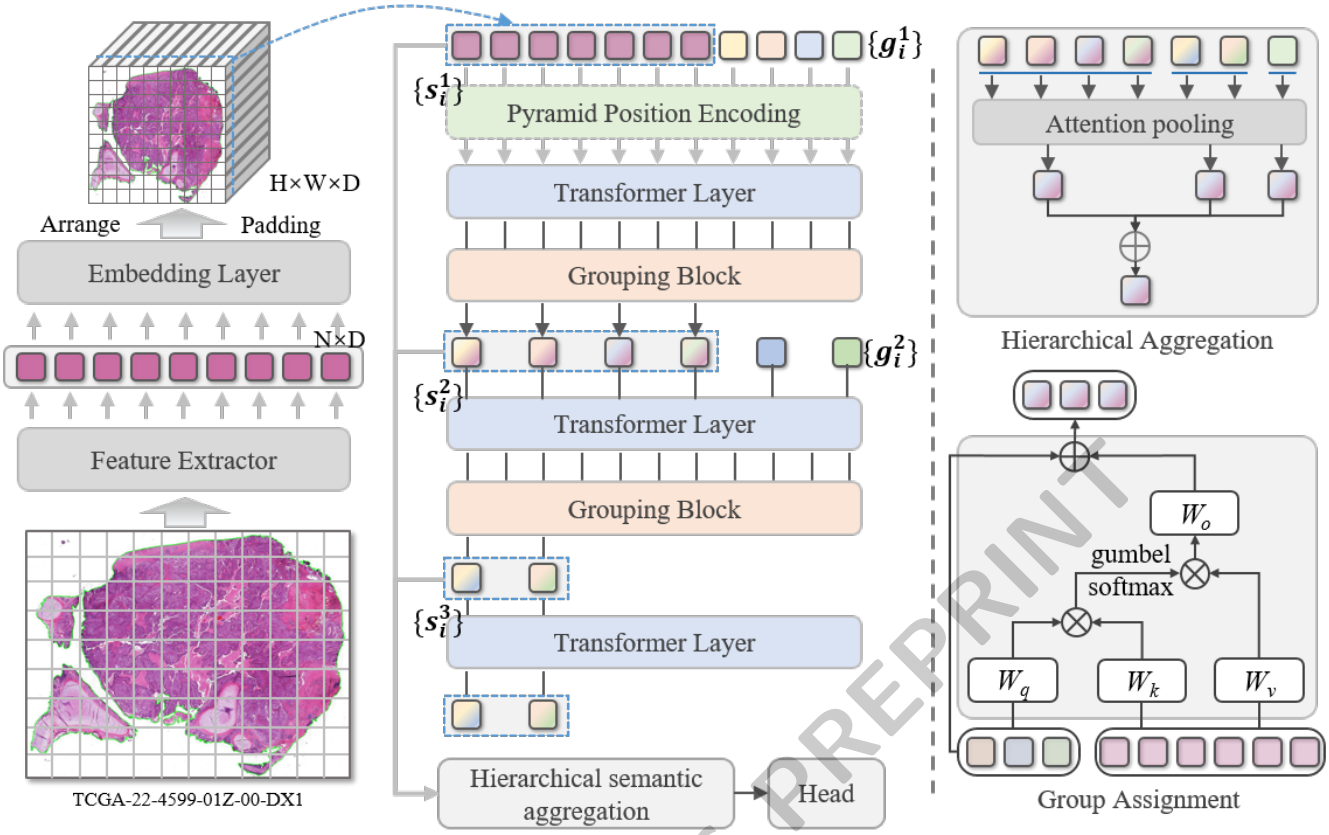

Figure 2: Schematic of the *GroupMIL* Architecture. The WSI is initially preprocessed by partitioning the tissue regions into patches, followed by the extraction of patch-level embeddings using a pre-trained feature extractor. These embeddings are subsequently processed through a multi-stage grouping, which adaptively cluster patches into semantic group tokens. Finally, the outputs from multiple grouping stages are fused via Hierarchical Semantic Aggregation to construct the slide-level representation.

a grouping layer. Through this hierarchical design, semantic segments within the WSI are progressively grouped. The detailed architecture is described bellow.

2.3 Adaptive Multi-Stage Grouping

WSIs are scanned at multiple magnification levels, and pathological tissues often exhibit a distinct hierarchical structure. Accordingly, *GroupMIL* adopts a multi-stage grouping operation to aggregate tissue information from fine to coarse granularity.

Formally, a WSI can be patched into N tokens after preprocessing following [Lu *et al.*, 2020], denoted as $\{p_i\}^N$, $p_i \subseteq \mathbb{R}^{1\times1024}$. Assuming the model has L stages, the input contains image tokens $\{s_i^l\}_{i=1}^{C_{l-1}}$ and group tokens $\{g_i^l\}_{i=1}^{C_l}$ at stage l, where $C_0 = N$. We simplify $\{s_i^l\}_{i=1}^{C_{l-1}}$ to $\{s_i^l\}$ and similarily $\{g_i^l\}_{i=1}^{C_l}$ to $\{g_i^l\}$. To enbale the information propagation between $\{s_i^l\}$ and $\{g_i^l\}$, we first concatenate them together and put them into a Transformer layer:

$$\{\hat{s}_i^l\}, \{\hat{g}_i^l\} = Transformer([\{s_i^l\}; \{g_i^l\}]) \tag{2}$$

where $[;]$ denotes the concatenate operator. Following that, the image tokens and group tokens are fed into a grouping block to obtain the new image tokens

$$\{s_i^{l+1}\} = GroupingBlock(\{\hat{s}_i^l\}, \{\hat{g}_i^l\}) \tag{3}$$

By merging multiple image tokens into a single new token, the $GroupingBlock$ facilitates a semantic hierarchical aggregation of the image tokens. Similar to [Jang *et al.*, 2016][Vaswani *et al.*, 2017], we compute the similarity matrix A^l between $\{\hat{s}_i^l\}$ and $\{\hat{g}_i^l\}$ via a $Gumble - Softmax$, as shown in Fig. 2. The operation is defined as

$$A_{i,j}^l = \frac{exp(W_q\hat{g}_i^l \cdot W_k\hat{s}_j^l + \gamma_i)}{\sum_{k=1}^{C_l} exp(W_q\hat{g}_k^l \cdot W_k\hat{s}_j^l + \gamma_k)} \tag{4}$$

where W_q and W_k are the learnable parameters of the Linear layer, and γ_i is an i.i.d sample drawn from the $Gumble(0, 1)$ distribution. We use the straight through and hard assignment trick in [Xu *et al.*, 2022][van den Oord *et al.*, 2017] to make the end to end training and the final output of $GroupingBlock$ is a weighted sum of image tokens assigned

$$s_i^{l+1} = \hat{g}_i^l + W_o \frac{\sum_{j=1}^{C_{l-1}} \hat{A}_{i,j}^l W_v \hat{s}_j^l}{\sum_{j=1}^{C_{l-1}} \hat{A}_{i,j}^l} \tag{5}$$

where W_v and W_o are the learnable parameters of the Linear layer. The final output is a set of grouped image tokens $\{s_i^l\}_{i=1}^{C_L}$.

2.4 Hierarchical Semantic Aggregation

Based on the multi-stage grouping described above, we perform hierarchical aggregation by integrating the outputs from all stages. At each stage, multiple output tokens are fused into a single token through an attention mechanism. The tokens obtained from all stages are then averaged to derive the final WSI representation.

Formally, the fused feature f^l is the weighted sum of $\{\hat{s}_i^l\}$:

$$f^l = \sum_{k=1}^{C_l} a_k^l \hat{s}_k^l \tag{6}$$

where

$$a_k^l = \frac{exp\{W_t^T tanh(W_v \hat{s}_k^{l\,T})\}}{\sum_{j=1}^{C_l} exp\{W_t^T tanh(W_v \hat{s}_j^{l\,T})\}} \tag{7}$$

where W_t and W_v are the learnable parameters of the Linear layer. In this way, multi-stage features are obtained and the WSI feature is derived by averaging these features

$$f = \frac{1}{L} \sum_{l=1}^{L} f^l \tag{8}$$

This slide-level representation is subsequently utilized for downstream clinical tasks, specifically cancer subtyping and survival prediction. For the subtyping task, we employ a fully-connected layer optimized via a cross-entropy loss. For survival analysis, we utilize a specialized output head optimized by a Cox proportional hazards loss [Katzman *et al.*, 2016], which accounts for censored survival data.

2.5 Long Patch Sequences Modeling and Position Embedding

Before *GroupMIL* can be effectively applied, an inevitable challenge must first be addressed: the computational complexity inherent in Transformer-based architecture. The self-attention computation among patch tokens and the similarity matrix calculation between patch tokens and group tokens suffer from $O\left(n^2\right)$ time and memory complexity. This significantly hampers the model's practical applicability. To deal with the long sequence modeling problems in WSIs, the standard row-by-row softmax function is replaced by the Nystrom Method proposed in [Xiong *et al.*, 2021]. The approximated attention form $\hat{S}$ can be defined as:

$$\hat{S} = softmax(\frac{Q\tilde{K}^T}{\sqrt{d^q}}) A_s^+ softmax(\frac{\tilde{Q}K^T}{\sqrt{d^q}}) \tag{9}$$

where $\tilde{Q}$ and $\tilde{K}$ are the m selected landmarks from the original n dimensional sequence of Q and K. A^+ is a Moore-Penrose pseudoinverse of a submatrix obtained by sampling the original attention matrix, with dimensions $m \times m$. When the number of landmarks m is much smaller than the sequence length n, the computational complexity is approximately reduced from $O(n^2)$ to $O(n)$. By doing this, *GroupMIL* becomes capable of processing WSI containing thousands of patches.

Following the architecture established by [Shao *et al.*, 2021], we incorporate Pyramid Positional Encoding to model the spatial relationships between patches. To handle the variable number of patches within a WSI, we apply zero-padding to the slide feature map, enabling the application of 2D convolutional neural networks. Specifically, we utilize three convolutional kernels with varying sizes 3, 5 and 7 to encode positional information at different levels of granularity. This multi-scale approach can encode the positional information with different granularity.

3 Experiments

To demonstrate the superior performance of the proposed *GroupMIL* model, we conducted extensive experiments on two classic tasks: WSI-based cancer subtype and prognosis prediction. For the cancer subtype task, validation was performed on 3 public challenging datasets: $\boldsymbol{BRACS}$, $\boldsymbol{NSCLC}$ and $\boldsymbol{TCGA\text{-}RCC}$. For the prognosis prediction task, validation was carried out on 5 pubilc datasets, including $\boldsymbol{TCGA\text{-}BRCA}$, $\boldsymbol{TCGA\text{-}KIRC}$, $\boldsymbol{TCGA\text{-}KIRP}$, $\boldsymbol{TCGA\text{-}LUAD}$ and $\boldsymbol{TCGA\text{-}STAD}$. All WSIs were preprocessed according to the protocol established in [Lu *et al.*, 2020], and all subsequent analyses were performed at a magnification of 20×. A detailed description of the experiments will be provided in the following sections.

3.1 Datasets and Evaluation Metrics

Cancer subtype

We perform comparative experiments on three public datasets: $\boldsymbol{BRACS}$, $\boldsymbol{NSCLC}$ and $\boldsymbol{TCGA\text{-}RCC}$. For the $\boldsymbol{BRACS}$ dataset[Brancati *et al.*, 2021], we perform experiments on the official split for fair comparison with existing methods. For the other two datasets, we employ 5-fold Monte Carlo cross-validaton, which partitions the data into training, validation, and testing sets with a ratio of 3:1:1. For the evaluation metrics, we used accuracy (ACC) and the Aera Under Curve (AUC) to evaluate the classification performance.

Prognosis prediction

We compare the performance of various models for tumor prognosis prediction across five public datasets. To ensure a fair comparison, we employed 5-fold cross-validation for each dataset, splitting the data into training and testing sets at a 4:1 ratio in each fold. The average Concordance Index (CI) from the cross-validation was used as the primary evaluation metric for model performance, and we calculated the standard deviation (std) to account for statistical variability.

3.2 Implementation Details

We present the experimental results of the proposed *GroupMIL* on several datasets, and compare it against the following methods: (1) conventional pooling methods, including Max-pooling and Mean-pooling; (2) i.i.d hyposis based methods, including ABMIL[Ilse *et al.*, 2018], CLAM[Lu *et al.*, 2020], DSMIL[Li *et al.*, 2020] and AttriMIL[Cai *et al.*, 2025]; (3) correlated MIL methods, including TransMIL[Shao *et al.*, 2021] and MambaMIL[Yang *et al.*, 2024]. We following the same data pre-processing as in the CLAM and set a learning rate of 5e-5 for all these model. The features of each patch are embedded in a 1024-dimensional vector by a ResNet50[He *et al.*, 2015] model pretrained on ImageNet[Deng *et al.*, 2009]. We incorporate a projection layer which reduces the dimensionality of the input embeddings from 1024 to 512. During

training, we employed an early stopping strategy to prevent overfitting and halt the training procedure when the validation loss ceased to decrease for a consecutive period of 20 epochs. All experiments were done on a server with 8 NVIDIA A6000 GPUs, utilizing the PyTorch framework.

3.3 Comparison Results

Cancer subtype

We will present the results of both binary and multi-class classification. The $\boldsymbol{NSCLC}$ involves a binary classification task distinguishing between LUSC and LUAD subtypes. $\boldsymbol{RCC}$ constitutes a three-class classification task differentiating among KICH, KIRC, and KIRP. $\boldsymbol{BRACS}$ encompasses a more complex classification task covering seven distinct breast lesion subtypes. All the experiment results are provided in Table 1.

As demonstrated in Table 1, our proposed *GroupMIL* model achieves performance improvements across different tasks compared to the previous state-of-the-art. It attains an AUC of 0.825 on the BRACS dataset, 0.956 on the NSCLC dataset, and 0.983 on the RCC dataset. These results demonstrate the superiority of the proposed model.

Prognosis prediction

In addition to cancer subtyping, we also conducted experiments on tumor prognosis prediction. We compared the proposed *GroupMIL* with eight other MIL methods across five public datasets, with the results presented in Table 2.

The experimental results in Table 2 demonstrate that our proposed *GroupMIL* model achieved the best performance across all datasets, confirming its effectiveness. Notably, on the five datasets, the *GroupMIL* model improved the CI by an average of 2% compared to the second-best performance model.

3.4 Ablation Study

To further analyze the impact of the model architecture on the experimental results, we conducted ablation studies to investigate the effects of the number of group centers and hierarchical aggregation on model performance. Specifically, ablation experiments were performed on the $\boldsymbol{NSCLC}$ dataset, with model performance evaluated using the AUC metric.

Effectiveness of Adaptive Multi-stage Grouping

The primary motivation for Adaptive multi-stage Grouping is to transition from fine-grained histological information to coarse-grained semantic regions. We first established a baseline model by removing the grouping mechanism, which effectively reduces the framework to a standard ABMIL[Ilse *et al.*, 2018] model.We then evaluated single-stage grouping by setting the number of group tokens to $K_1 = 16$ and $K_1 = 32$, respectively. We compared the AUC under different model, with the results summarized in Table 3.

The results demonstrate that while a single stage of grouping improves over the i.i.d. baseline by introducing semantic structure, the Multi-stage approach provides a better performance. This suggests that the hierarchical structure leading to a more robust slide-level representation.

Impact of Semantic Group token Number

We further investigated the impact of group token cardinality at each stage on model performance. Specifically, we varied the number of tokens across the multi-stage architecture to determine the optimal balance between feature effectiveness and semantic representation. The experiment results are detailed in Table 4.

We observed that setting K too large may introduces redundant semantic features, which lead to overfitting on datasets. Conversely, setting K too small may results in a over simplification, where distinct tissue types (e.g., tumor and stroma) are forced into the same group. The combination of 32 fine-grained and 16 coarse-grained tokens provides the optimal granularity for capturing pathological diversity.

3.5 Interpretability and Visualization

By assigning each patch embedding to its corresponding semantic group token, *GroupMIL* inherently partitions the WSI into distinct clusters. This grouping mechanism allows us for a direct visualization of the semantic distribution across the slide. To interpret these results, we projected the discrete group assignments back onto the original WSI coordinates.

As illustrated in Fig. 3, the model demonstrates a high degree of morphological consistency. Patches with similar histological characteristics are reliably clustered into same group tokens. This alignment suggests that the learned semantic group tokens effectively captures the underlying biological structures of the tissue, enabling it to distinguish between critical phenotypes, such as tumor regions and surrounding stroma.

4 Conclusion

This study proposed *GroupMIL*, a novel MIL framework that differs from the traditional i.i.d. based MIL or globally correlation MIL to aggregate the WSI information. By leveraging the Adaptive Multi-stage Grouping module, *GroupMIL* effectively captures the latent histological organization of WSIs, transforming thousands of patches into representative tissue phenotypes. Furthermore, the Hierarchical Semantic Aggregation module ensures that both fine-grained morphological details and coarse-grained structural contexts are preserved and integrated into the final slide-level representation. Extensive experiments across multiple public datasets demonstrate that *GroupMIL* consistently achieve satisfactory performance in cancer subtyping and survival prediction. The grouping mechanism also provides a clinical interpretability, enabling the automated discovery of diagnostically significant regions without the need for dense pixel-level annotations. Future work will explore the integration of Large Language Models (LLMs) to provide high-level semantic guidance, utilizing medical knowledge to further refine the patch-based groups.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62501232), the Science and Technology In-

Model	BRACS-7		NSCLC		TCGA-RCC		MEAN	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
Max-pooling	0.687	0.395	$0.930_{\pm 0.020}$	$0.860_{\pm 0.025}$	$0.973_{\pm 0.008}$	$0.889_{\pm 0.018}$	0.863	0.715
Mean-pooling	0.639	0.301	$0.871_{\pm 0.020}$	$0.779_{\pm 0.017}$	$0.950_{\pm 0.012}$	$0.822_{\pm 0.029}$	0.820	0.634
ABMIL[Ilse *et al.*, 2018]	0.793	0.395	$0.915_{\pm 0.017}$	$0.856_{\pm 0.024}$	$0.925_{\pm 0.018}$	$0.806_{\pm 0.025}$	0.878	0.686
CLAM-MB[Lu *et al.*, 2020]	0.779	0.382	$0.912_{\pm 0.026}$	$0.827_{\pm 0.041}$	$0.979_{\pm 0.013}$	$0.891_{\pm 0.046}$	0.890	0.700
DSMIL[Li *et al.*, 2020]	0.736	0.335	$0.792_{\pm 0.034}$	$0.728_{\pm 0.039}$	$0.957_{\pm 0.020}$	$0.838_{\pm 0.047}$	0.828	0.634
TransMIL[Shao *et al.*, 2021]	0.765	0.349	$0.927_{\pm 0.019}$	$0.870_{\pm 0.029}$	$0.976_{\pm 0.010}$	$0.911_{\pm 0.030}$	0.889	0.710
MambaMIL[Yang *et al.*, 2024]	0.752	0.361	$0.922_{\pm 0.016}$	$0.847_{\pm 0.027}$	$0.945_{\pm 0.011}$	$0.901_{\pm 0.029}$	0.873	0.703
AttriMIL[Cai *et al.*, 2025]	0.689	0.372	$0.928_{\pm 0.010}$	$0.849_{\pm 0.018}$	$0.975_{\pm 0.017}$	$0.913_{\pm 0.028}$	0.864	0.711
GroupMIL(Ours)	**0.825**	**0.512**	**$0.956_{\pm 0.010}$**	**$0.893_{\pm 0.027}$**	**$0.983_{\pm 0.012}$**	**$0.920_{\pm 0.031}$**	**0.921**	**0.775**

Table 1: Cancer subtyping results of difference models on $\boldsymbol{BRACS}$, $\boldsymbol{NSCLC}$ and $\boldsymbol{TCGA\text{-}RCC}$.

Model	BRCA	LUAD	STAD	KIRC	KIRP	MEAN
Max-pooling	$0.546_{\pm 0.092}$	$0.607_{\pm 0.052}$	$0.539_{\pm 0.046}$	$0.604_{\pm 0.035}$	$0.646_{\pm 0.092}$	0.588
Mean-pooling	$0.572_{\pm 0.052}$	$0.609_{\pm 0.042}$	$0.592_{\pm 0.047}$	$0.644_{\pm 0.019}$	$0.673_{\pm 0.019}$	0.618
ABMIL[Ilse *et al.*, 2018]	$0.601_{\pm 0.020}$	$0.609_{\pm 0.053}$	$0.584_{\pm 0.062}$	$0.641_{\pm 0.040}$	$0.717_{\pm 0.062}$	0.630
CLAM-MB[Lu *et al.*, 2020]	$0.653_{\pm 0.062}$	$0.629_{\pm 0.048}$	$0.633_{\pm 0.036}$	$0.725_{\pm 0.042}$	$0.709_{\pm 0.037}$	0.670
DSMIL[Li *et al.*, 2020]	$0.598_{\pm 0.046}$	$0.615_{\pm 0.052}$	$0.638_{\pm 0.038}$	$0.649_{\pm 0.033}$	$0.707_{\pm 0.052}$	0.641
TransMIL[Shao *et al.*, 2021]	$0.602_{\pm 0.037}$	$0.636_{\pm 0.049}$	**$0.665_{\pm 0.042}$**	$0.645_{\pm 0.039}$	$0.658_{\pm 0.097}$	0.641
MambaMIL[Yang *et al.*, 2024]	$0.641_{\pm 0.028}$	$0.628_{\pm 0.069}$	$0.639_{\pm 0.036}$	$0.721_{\pm 0.079}$	$0.708_{\pm 0.047}$	0.667
AttriMIL[Cai *et al.*, 2025]	$0.631_{\pm 0.046}$	$0.622_{\pm 0.047}$	$0.641_{\pm 0.052}$	$0.684_{\pm 0.058}$	$0.701_{\pm 0.062}$	0.656
GroupMIL(Ours)	**$0.662_{\pm 0.062}$**	**$0.654_{\pm 0.023}$**	$0.650_{\pm 0.045}$	**$0.749_{\pm 0.028}$**	**$0.737_{\pm 0.058}$**	**0.690**

Table 2: Survival prediction of difference models on five datasets.

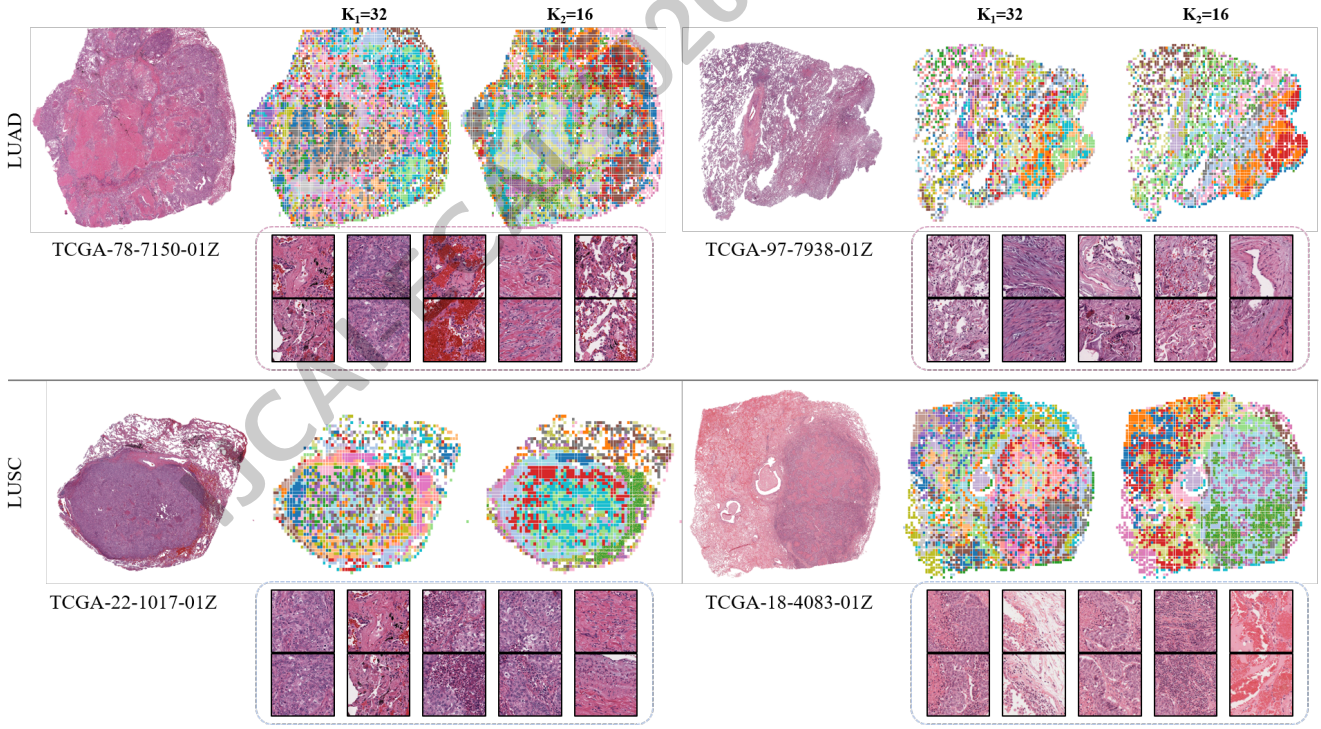

Figure 3: Visualization of Semantic Groups. Examples from the LUAD and LUSC datasets illustrate the grouping results under configurations of $K_1 = 32$ and $K_2 = 16$ group tokens, respectively. For each configuration, two randomly selected WSIs are displayed alongside two sample patches extracted from the same semantic group. The visualization demonstrates that *GroupMIL* achieves high morphological consistency, effectively clustering patches with similar pattern into semantic units.

Configuration	Grouping K_1	Grouping K_2	AUC
Baseline	-	-	0.915
Single-Stage	K_1=16	-	0.923
Single-Stage	K_1=32	-	0.931
Multi-Stage	K_1=32	K_2=16	**0.956**

Table 3: Effects of Adaptive Multi-stage Grouping

Stage1(K_1)	Stage2(K_2)	AUC
64	32	0.935
32	32	0.929
32	16	**0.956**
16	8	0.907

Table 4: Impact of Group Token Number

novation Program of Hunan Province (2025RC3098), the National Natural Science Foundation of China (62425305), the Research Foundation of Education Bureau of Hunan Province (24B0037) and the Fundamental Research Funds for the Central Universities (531118010979).

References

[Brancati *et al.*, 2021] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniele Riccio, G. Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database: The Journal of Biological Databases and Curation*, 2022, 2021.

[Cai *et al.*, 2025] 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.

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

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[He *et al.*, 2015] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[Ilse *et al.*, 2018] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. *ArXiv*, abs/1802.04712, 2018.

[Jang *et al.*, 2016] Eric Jang, Shixiang Shane Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *ArXiv*, abs/1611.01144, 2016.

[Kanavati *et al.*, 2020] Fahdi Kanavati, Gouji Toyokawa, Seiya Momosaki, Michael Rambeau, Yuka Kozuma, Fumihiro Shoji, Koji Yamazaki, Sadanori Takeo, Osamu Iizuka, and Masayuki Tsuneki. Weakly-supervised learning for lung carcinoma classification using deep learning. *Scientific Reports*, 10, 2020.

[Katzman *et al.*, 2016] Jared Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18, 2016.

[Li *et al.*, 2020] Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323, 2020.

[Lu *et al.*, 2020] 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, 2020.

[Niazi *et al.*, 2019] Muhammad Khalid Khan Niazi, Anil V. Parwani, and Metin N. Gurcan. Digital pathology and artificial intelligence. *The Lancet. Oncology*, 20 5:e253–e261, 2019.

[Ning *et al.*, 2025] Zehang Ning, Bojie Yang, Yuanyuan Wang, Zhifeng Shi, Jinhua Yu, and Guoqing Wu. Dual-path neural network extracts tumor microenvironment information from whole slide images to predict molecular typing and prognosis of glioma. *Computer methods and programs in biomedicine*, 261:108580, 2025.

[Pati *et al.*, 2023] Pushpak Pati, Guillaume Jaume, Zeineb Ayadi, Kevin Thandiackal, Behzad Bozorgtabar, Maria Gabrani, and Orcun Goksel. Weakly supervised joint whole-slide segmentation and classification in prostate cancer. *Medical image analysis*, 89:102915, 2023.

[Sakaguchi *et al.*, 2025] Maki Sakaguchi, Akihiko Yoshizawa, Kenta Masui, Tomoya Sakai, and Takashi Komori. Ai-powered histology for molecular profiling in brain tumors: Toward smart diagnostics from tissue. *Cancers*, 2025.

[Shao *et al.*, 2021] Zhucheng 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 classication. In *Neural Information Processing Systems*, 2021.

[Tian *et al.*, 2024] Mengxin Tian, Zhao Yao, Yufu Zhou, Qiangjun Gan, Leihao Wang, Hongwei Lu, Siyuan Wang, Peng Zhou, Zhiqiang Dai, Sijia Zhang, Yihong Sun, Zhaoqing Tang, Jinhua Yu, and Xuefei Wang. Deeprisk network: an ai-based tool for digital pathology signature and treatment responsiveness of gastric cancer using whole-slide images. *Journal of Translational Medicine*, 22, 2024.

[van den Oord *et al.*, 2017] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Neural Information Processing Systems*, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[Wang *et al.*, 2016] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognit.*, 74:15–24, 2016.

[Xiong *et al.*, 2021] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Moo Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 35 16:14138–14148, 2021.

[Xu *et al.*, 2022] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and X. Wang. Groupvit: Semantic segmentation emerges from text supervision. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18113–18123, 2022.

[Yang *et al.*, 2024] Shu Yang, Yihui Wang, and Hao Chen. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[Zhang *et al.*, 2023a] Ruijie Zhang, Qiaozheng Zhang, Yingzhuang Liu, Hao Xin, Y. Liu, and Xinggang Wang. Multi-level multiple instance learning with transformer for whole slide image classification. *ArXiv*, abs/2306.05029, 2023.

[Zhang *et al.*, 2023b] Yunlong Zhang, Honglin Li, Yuxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. *ArXiv*, abs/2311.07125, 2023.