AceMIL: Ordinal-Aware Multiple Instance Learning for Pathological Progression Analysis

Supplementary Material

Table 7. **Evaluation on NAS-Steatosis.** Metrics are averaged over 5-fold cross-validation. κ denotes quadratic-weighted kappa. AUC and F1 scores are micro-averaged.

Methods	κ	Accuracy	AUC	F1
ABMIL	0.5175	0.5	0.6194	0.4718
DSMIL	0.8315	0.7370	0.9020	0.7335
CLAM	0.8695	0.7888	0.9445	0.7896
TransMIL	0.8597	0.7722	0.9303	0.7701
MambaMIL	0.8597	0.7796	0.9422	0.7757
WIKG	0.8732	0.7962	0.9456	0.7949
AceMIL	**0.8938**	**0.8351**	**0.9548**	**0.8360**

Table 8. **Evaluation on NAS-Inflammation.** Metrics are averaged over 5-fold cross-validation. κ denotes quadratic-weighted kappa. AUC and F1 scores are micro-averaged.

Methods	κ	Accuracy	AUC	F1
ABMIL	0.4847	0.6593	0.7631	0.6220
DSMIL	0.5289	0.6370	0.8197	0.6087
CLAM	0.5365	0.6463	0.8402	0.6130
TransMIL	0.4684	0.6037	0.7797	0.5742
MambaMIL	0.5674	0.6519	0.8415	0.6316
WIKG	0.6141	0.6648	**0.8554**	0.6407
AceMIL	**0.6715**	**0.6741**	0.8207	**0.6633**

Table 9. **Evaluation on NAS-Ballooning.** Metrics are averaged over 5-fold cross-validation. κ denotes quadratic-weighted kappa. AUC and F1 scores are micro-averaged.

Methods	κ	Accuracy	AUC	F1
ABMIL	0.5001	0.7278	0.7316	0.7093
DSMIL	0.6002	0.7648	0.8696	0.7549
CLAM	0.6176	0.7704	**0.8753**	0.7642
TransMIL	0.5471	0.7259	0.8248	0.7210
MambaMIL	0.6246	0.7722	0.8713	0.7715
WIKG	0.6177	0.7630	0.8734	0.7614
AceMIL	**0.6641**	**0.7870**	0.8512	**0.7812**

6. Supplementary Experiments

Performance comparison on sub-tasks of NAS: Steatosis, Inflammation, and Ballooning.

As we depict in Section 4.1, NAS-8s is composed of three sub-tasks, but we directly deploy eight-stage modeling. Here, we split it into three sub-tasks and give the comparison of AceMIL with other SOTA methods separately. Shown in Table 7-9, AceMIL still acquires the best accuracy and κ, demonstrating its generalization.

Table 10. Efficiency comparison against SOTA methods. The meaning of † is provided in Table 5.

Methods	Param.	GFLOPs†	Throughput
ABMIL	1.53M	4.56	1141 WSIs / s
DSMIL	1.26M	3.53	873 WSIs / s
CLAM	1.82M	5.38	574 WSIs / s
TransMIL	8.96M	59.16	62 WSIs / s
MambaMIL	14.89M	39.46	86 WSIs / s
WIKG	0.57M	19.65	43 WSIs / s
AceMIL	1.94M	12.06	145 WSIs / s

Efficiency evaluation.

We compare the efficiency of AceMIL with representative MIL-based methods in terms of model size, computational cost (GFLOPs†), and throughput (WSIs/s), as summarized in Table 10. Overall, AceMIL achieves a favorable trade-off between performance and efficiency: it processes 145 WSIs per second while maintaining a moderate model size (1.94M parameters) and computational cost (12.06 GFLOPs). Although classical MIL methods like ABMIL, DSMIL, and CLAM report higher throughput, they exhibit significantly lower model capacity and inferior performance. Compared to advanced methods such as TransMIL (62 WSIs/s), MambaMIL (86 WSIs/s), and WIKG (43 WSIs/s), AceMIL offers substantially better throughput with fewer parameters and lower computation. These results demonstrate that AceMIL achieves superior efficiency without compromising accuracy, making it well-suited for real-world clinical deployment.

Visualization of soft labels on Camelyon16.

The annealing soft label of NAS-8s has been illustrated in text. We supplement the visualization of the annealing soft labels on Camelyon17 for reference just as shown in Figure 6.

Patch-level significance and pseudo-stage visualization of the PSA.

Patch-level Stage Aggregation (PSA) is capable of predicting both the significance and the pathological stage of individual patches. To validate our motivation and provide interpretable insights, we present attention maps and stage visualizations on WSIs. As shown in Figure 7, several key

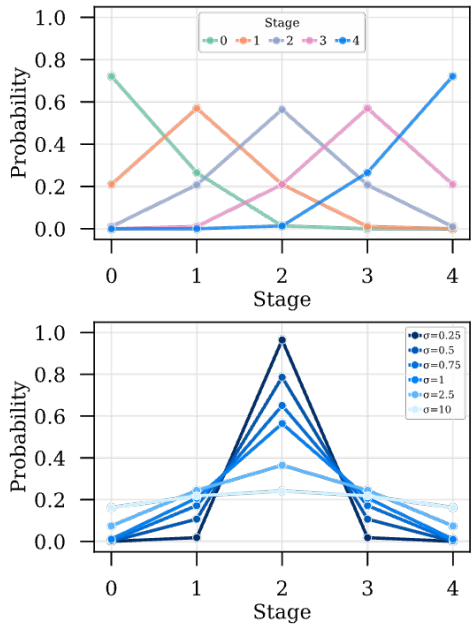

Figure 6. **Soft label distribution of Camelyon17. Up**: Smoothed stage distribution from 0 to 4 while temperature $\sigma = 1$. **Bottom**: As the annealing temperature (σ from 10 to 0.25) decrease, stage distribution gradually sharpens from uncertainty to confidence.

observations can be made. First, benefiting from the salient modeling capability of the DIFF Transformer, the attention mechanism focuses sparsely on the most informative patches. Second, the predicted patch-level pseudo-stages offer diverse local ordinal cues; however, some noisy or inconsistent predictions are observed due to the lack of direct supervision. Alternatively, it is also possible that PSA is not learning true ordinal stages, but instead discovering latent clusters among patches. Despite this, the potential for clinical translation remains promising, especially when integrated with physician expertise. Overall, these findings demonstrate that PSA effectively captures patch-level significance and local ordinal information, aligning well with our original motivation for proposing PSA.

7. Data and Implementation Details

NAS-8s. The NAS-8s dataset is derived from non-alcoholic fatty liver disease (NAFLD) patients and represents the NAFLD Activity Score (NAS), where "8s" denotes that the total score ranges from 0 to 8. NAS comprises three histological sub-staging components: steatosis, inflammation, and ballooning, independently graded from 0–4, 0–4, and 0–3, respectively. These morphological features are strongly associated with non-alcoholic steatohepatitis (NASH). The sum of the three sub-scores constitutes the final NAS stage. We primarily conduct experiments on the overall 9-stage NAS prediction task and report additional results for the three sub-staging tasks in the supplement. The dataset includes 542 hematoxylin and eosin (H&E)-stained whole-slide images (WSIs), each obtained from a unique patient and labeled by consensus among three experienced pathologists in accordance with the NASH-CRN system [21]. Data release is subject to approval from the institutional medical ethics committee.

Camelyon17 pN-stage. The Camelyon17 dataset [4] contains 1,000 H&E-stained lymph node WSIs collected from 200 breast cancer patients, with five slides per patient. It focuses on the detection and classification of breast cancer metastases in axillary lymph nodes. Slides are evaluated using the widely adopted TNM staging system, with specific emphasis on the pathologic N-stage (pN-stage), which reflects the extent of lymph node metastasis. The pN-stage is determined based on the number and size of metastatic-positive lymph nodes, and includes five ordinal categories: pN0 (no metastasis), pN0(i+) (isolated tumor cells), pN1mi (micro-metastases), pN1, and pN2, representing increasing levels of disease progression. CAMELYON17's pN labels represent a clinically validated, monotonic progression of nodal metastatic burden rather than interval-scale measurements; accordingly, treating pN as an ordinal target is both clinically justified and consistent with the CAMELYON17 benchmark protocol and prior literature.

TUPAC16. The TUPAC16 dataset [52] comprises H&E-stained breast cancer whole-slide images (WSIs) collected for the Tumor Proliferation Assessment Challenge 2016, containing 500 training and 321 testing slides. While only the labels of the 500 training samples are publicly available, we solely utilize them for benchmarking. TUPAC16 focuses on predicting tumor proliferation from WSIs through two complementary tasks: the pathologist-assigned mitotic score and a gene-expression-based proliferation score. The mitotic score, ranging from 1 to 3, represents an ordinal scale reflecting increasing proliferative activity and tumor aggressiveness, while the continuous gene-expression score provides a quasi-ordinal measure of proliferation. We use the pathologist-assigned mitotic score as the target. These scores capture a monotonic progression of tumor proliferation rather than interval-scale differences. Therefore, modeling TUPAC16 scores as ordinal targets is both clinically meaningful and consistent with our established evaluation protocols.

Implementation details. The model is trained using the AdamW [32] optimizer with a learning rate of 2×10^{-4}, a batch size of 1, and a total of 100 epochs. For other MIL methods, standard CrossEntropy loss is adopted. For ordinal supervision, we adopt the proposed Annealing Soft Label (ASL) framework using KL divergence as the training loss. The initial temperature is set to $\sigma_{\text{start}} = 10$, and we conduct experiments with various $\sigma_{\text{end}} \in \{0.25, 0.5, 1.0, 2.5\}$ to assess the effect and convenience of

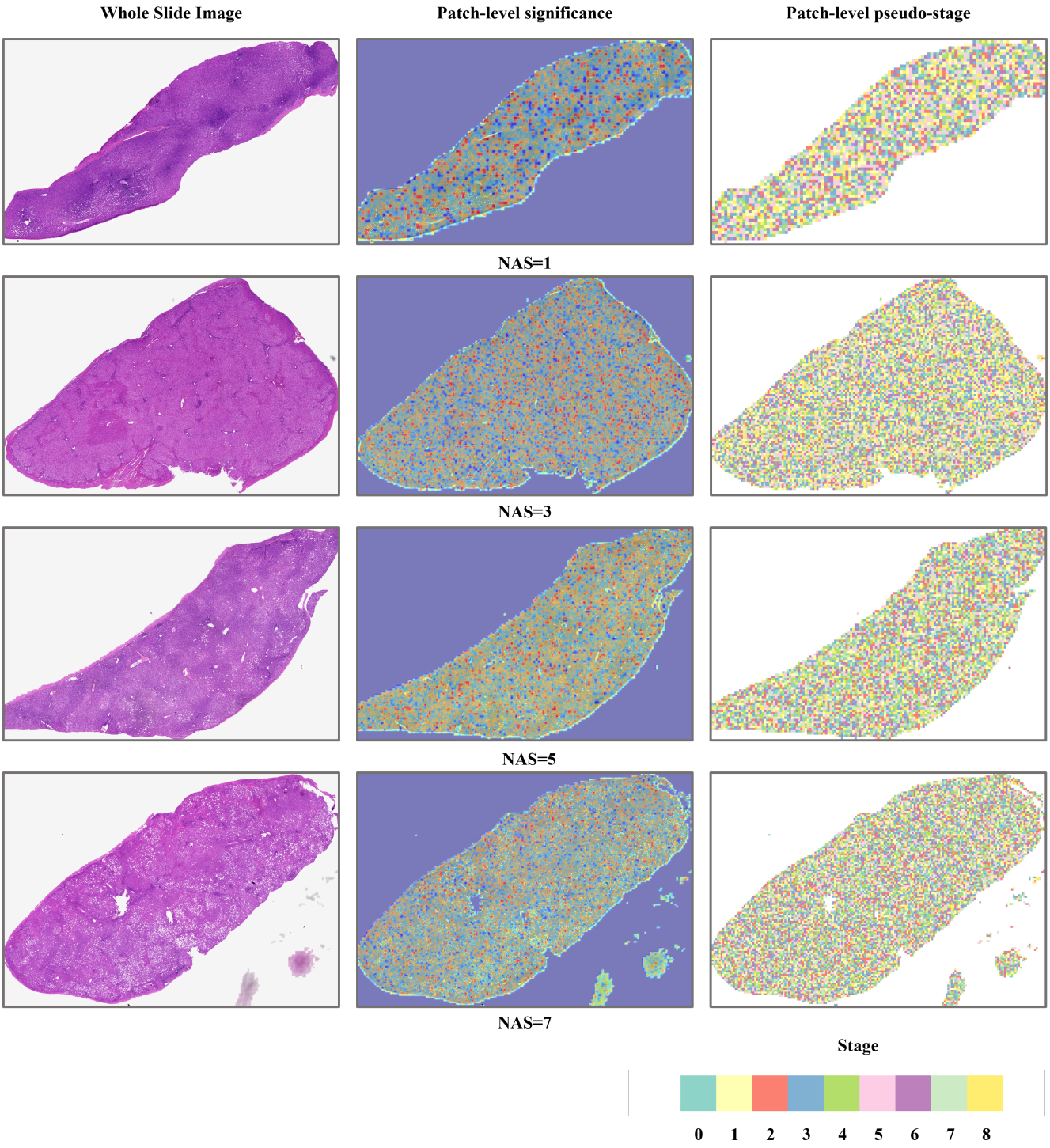

Figure 7. **Attention map and patch-level pseudo-stage of PSA visualization on NAS-8s.** Min-max normalization is adopted on significant values.

annealing soft label.