MExD: An Expert-Infused Diffusion Model for Whole-Slide Image Classification –Supplementary Material–

Jianwei Zhao[1], Xin Li[2], Fan Yang[2*], Qiang Zhai[3], Ao Luo[4], Yang Zhao[1], Hong Cheng[1], and Huazhu Fu[5]
[1]UESTC, [2]AIQ, [3]SICAU, [4]SWJTU, [5]IHPC, A*STAR

Abstract

This supplementary document offers an in-depth exploration of the methodologies, theoretical foundations, and experimental results presented in the main manuscript, providing additional analysis and valuable insights. To ensure a thorough understanding of our contributions, we include detailed derivations of the core equations underpinning the proposed approach, alongside a comprehensive exposition of the Diffusion Classifier architecture, emphasizing its design rationale and operational intricacies.

Furthermore, we present an expanded discussion on uncertainty quantification, highlighting its critical role in improving model reliability and interpretability. To substantiate the robustness and versatility of our method, we provide additional qualitative visualizations that demonstrate its efficacy across diverse scenarios. Lastly, we include an expert load analysis, illustrating how the model efficiently allocates computational resources and adapts dynamically to varying data complexities.

By addressing these elements, this supplementary material seeks to deepen the reader's understanding of our work and reinforce its relevance in advancing the state of the art.

1. Derivations

Forward Diffusion Sampling $\mathbf{f}_t$ at Arbitrary t steps. We incorporate the prior prediction, $\boldsymbol{\rho}_\theta$, from the Dyn-MoE aggregator into the forward diffusion process. In this section, we derive the parameters of the sampling distribution for the forward diffusion process at an arbitrary time step t. Following the methodology outlined in [7], the detailed formulation of Eq. 6 from the main manuscript is obtained through reparameterization techniques:

$$\begin{aligned}
\mathbf{f}_t &= \sqrt{\alpha_t}\mathbf{f}_{t-1} + (1-\sqrt{\alpha_t})\boldsymbol{\rho}_\theta + \sqrt{\beta_t}\epsilon && (1)\\
&= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}\mathbf{f}_{t-2} + (1-\sqrt{\alpha_{t-1}})\boldsymbol{\rho}_\theta + \sqrt{\beta_{t-1}}\epsilon_1) \\
&\quad + (1-\sqrt{\alpha_t})\boldsymbol{\rho}_\theta + \sqrt{\beta_t}\epsilon && (2)\\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{f}_{t-2} + (1-\sqrt{\alpha_t\alpha_{t-1}})\boldsymbol{\rho}_\theta \\
&\quad + \sqrt{\alpha_t}\sqrt{1-\alpha_{t-1}}\epsilon_1 + \sqrt{1-\alpha_t}\epsilon && (3)\\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{f}_{t-2} + (1-\sqrt{\alpha_t\alpha_{t-1}})\boldsymbol{\rho}_\theta \\
&\quad + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\epsilon}_1 && (4)\\
&\quad \cdots \\
&= \sqrt{\prod_{i=2}^{t}\alpha_i}\,\mathbf{f}_1 + \left(1-\sqrt{\prod_{i=2}^{t}\alpha_i}\right)\boldsymbol{\rho}_\theta \\
&\quad + \left(\sqrt{1-\prod_{i=2}^{t}\alpha_i}\right)\bar{\epsilon}_{t-2} && (5)\\
&= \sqrt{\prod_{i=2}^{t}\alpha_i}\,(\sqrt{\alpha_1}\mathbf{f}_0 + (1-\sqrt{\alpha_1})\boldsymbol{\rho}_\theta + \sqrt{\beta_1}\epsilon_{t-1}) \\
&\quad + \left(1-\sqrt{\prod_{i=2}^{t}\alpha_i}\right)\boldsymbol{\rho}_\theta + \left(\sqrt{1-\prod_{i=2}^{t}\alpha_i}\right)\bar{\epsilon}_{t-2} && (6)\\
&= \sqrt{\prod_{i=1}^{t}\alpha_i}\,\mathbf{f}_0 + \left(1-\sqrt{\prod_{i=1}^{t}\alpha_i}\right)\boldsymbol{\rho}_\theta \\
&\quad + \left(\sqrt{1-\prod_{i=1}^{t}\alpha_i}\right)\bar{\epsilon}_{t-1} && (7)\\
&= \sqrt{\bar{\alpha}_t}\mathbf{f}_0 + (1-\sqrt{\bar{\alpha}_t})\boldsymbol{\rho}_\theta + \sqrt{1-\bar{\alpha}_t}\bar{\epsilon}_{t-1}, && (8)
\end{aligned}$$

where $\alpha_t = 1-\beta_t$. As t increasing to T, $\sqrt{\bar{\alpha}_t}$ is limited to 0, where:

$$\mathbf{f}_T = \boldsymbol{\rho}_\theta + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \boldsymbol{I}). \tag{9}$$

Considering the independent additivity property of the Gaussian distribution, we observe that the endpoint of the forward diffusion process intuitively converges to $\boldsymbol{\rho}_\theta$, which can be formally defined as:

$$\mathbf{f}_T \sim \mathcal{N}(\boldsymbol{\rho}_\theta, \boldsymbol{I}). \tag{10}$$

*Corresponding author: Fan Yang

Posterior Mean $\widetilde{\boldsymbol{\mu}}$ and Variance $\widetilde{\boldsymbol{\beta}}$ in Reverse Denoising. Meanwhile, we integrated the prior prediction $\boldsymbol{\rho}_\theta$ into the reverse denoising process, significantly enhancing its efficiency. In this section, we derive the mean $\widetilde{\boldsymbol{\mu}}$ and variance $\widetilde{\boldsymbol{\beta}}$ for the reverse denoising process, as formalized in Eq. 7 of the main manuscript. Utilizing conditional Bayesian principles, the reverse denoising process can be decomposed into a series of sub-processes, which can be formulated as:

$$q(\mathbf{f}_{t-1}|\mathbf{f}_t,\mathbf{f}_0,\boldsymbol{\rho}_\theta)=\frac{q(\mathbf{f}_t|\mathbf{f}_{t-1},\boldsymbol{\rho}_\theta)q(\mathbf{f}_{t-1}|\mathbf{f}_0,\boldsymbol{\rho}_\theta)}{q(\mathbf{f}_t|\mathbf{f}_0,\boldsymbol{\rho}_\theta)} \tag{11}$$

$$\propto \exp\Big(-\frac{1}{2}\Big(\frac{\big(\mathbf{f}_t-(\sqrt{\alpha_t}\mathbf{f}_{t-1}+(1-\sqrt{\alpha_t})\boldsymbol{\rho}_\theta)\big)^2}{\beta_t} + \frac{\big(\mathbf{f}_{t-1}-(\sqrt{\overline{\alpha}_{t-1}}\mathbf{f}_0+(1-\sqrt{\overline{\alpha}_{t-1}})\boldsymbol{\rho}_\theta)\big)^2}{1-\overline{\alpha}_{t-1}}\Big)\Big) \tag{12}$$

$$\propto \exp\Big(-\frac{1}{2}\Big(\frac{\alpha_t\mathbf{f}_{t-1}^2-2\sqrt{\alpha_t}\big(\mathbf{f}_t-(1-\sqrt{\alpha_t})\boldsymbol{\rho}_\theta\big)\mathbf{f}_{t-1}}{\beta_t} + \frac{\mathbf{f}_{t-1}^2-2\big(\sqrt{\overline{\alpha}_{t-1}}\mathbf{f}_0+(1-\sqrt{\overline{\alpha}_{t-1}})\boldsymbol{\rho}_\theta\big)\mathbf{f}_{t-1}}{1-\overline{\alpha}_{t-1}}\Big)\Big) \tag{13}$$

$$= \exp\Big(-\frac{1}{2}\Big(\underbrace{\big(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\overline{\alpha}_{t-1}}\big)}_{\boxed{1}}\mathbf{f}_{t-1}^2-2\underbrace{\big(\frac{\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_{t-1}}\mathbf{f}_0}_{\boxed{1}} + \underbrace{\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{f}_t+\big(\frac{\sqrt{\alpha_t}(\sqrt{\alpha_t}-1)}{\beta_t}+\frac{1-\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_{t-1}}\big)\boldsymbol{\rho}_\theta\big)\mathbf{f}_{t-1}}_{\boxed{2}} + C\big(\mathbf{f}_t,\mathbf{f}_0,\boldsymbol{\rho}_\theta\big)\Big)\Big), \tag{14}$$

Subsequently, taking standard Gaussian probability density function in Eq. 15 for reference:

$$\exp(-\frac{1}{2}(\frac{1}{\sigma^2}x^2-\frac{2\mu}{\sigma^2}x+\frac{\mu^2}{\sigma^2})), \tag{15}$$

We derive the posterior variance $\widetilde{\boldsymbol{\beta}}_t$, and $\gamma_3=\sqrt{\widetilde{\boldsymbol{\beta}}_t}$:

$$\widetilde{\boldsymbol{\beta}}_t=\frac{1}{\boxed{1}}=\frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\boldsymbol{\beta}_t, \tag{16}$$

Furthermore, we calculate the following coefficients γ_0,γ_1 and γ_2 in the posterior mean, formulated as:

$$\gamma_0=\frac{\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_{t-1}}/\boxed{1}=\frac{\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}\beta_t, \tag{17}$$

$$\gamma_1=\frac{\sqrt{\alpha_t}}{\beta_t}/\boxed{1}=\frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\sqrt{\alpha_t}, \tag{18}$$

$$\gamma_2=\Big(\frac{\sqrt{\alpha_t}(\sqrt{\alpha_t}-1)}{\beta_t}+\frac{1-\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_{t-1}}\Big)/\boxed{1} \tag{19}$$

$$=\frac{\alpha_t-\overline{\alpha}_t-\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})+\beta_t(1-\sqrt{\overline{\alpha}_{t-1}})}{1-\overline{\alpha}_t} \tag{20}$$

Table 1. **Architecture** of our Diff-C.

Input	$\boldsymbol{g}_\alpha$, $\mathbf{f}_t$, $\boldsymbol{\rho}_\theta$, t
Condition Encoder	$\boldsymbol{c}=\sum_{r=0}^{k}(\boldsymbol{c}_r\cdot\boldsymbol{e}_r)$
	$h_1=\varrho\,(f_{1,c}(\boldsymbol{c}))$
	$h_2=\varrho\,(f_{2,c}(h_1))$
	$Z=f_{3,c}(h_2)$
Noise Estimation	$h_{1,y}=\varrho\,(f_{1,y}(\mathbf{f}_t\oplus\boldsymbol{\rho}_\theta)\odot f_{1,t}(t))$
	$h_{1,t}=Z\odot h_{1,y}$
	$h_{2,t}=\varrho\,(f_{2,h}(h_{1,t})\odot f_{2,t}(t))$
	$h_{3,t}=\varrho\,(f_{3,h}(h_{2,t})\odot f_{3,t}(t))$
Output	$\epsilon_\theta=f_4(h_{3,t})$

$$=1+\frac{(\sqrt{\overline{\alpha}_t}-1)(\sqrt{\alpha_t}+\sqrt{\overline{\alpha}_{t-1}})}{1-\overline{\alpha}_t}. \tag{21}$$

Hence, the posterior mean $\widetilde{\boldsymbol{\mu}}$ in Eq. 7 (main manuscript) can be calculated as:

$$\widetilde{\boldsymbol{\mu}}(\mathbf{f}_t,\mathbf{f}_0,\boldsymbol{\rho}_\theta)=\gamma_0\mathbf{f}_0+\gamma_1\mathbf{f}_t+\gamma_2\boldsymbol{\rho}_\theta. \tag{22}$$

2. Diffusion Classifier (Diff-C) Architecture

We present the detailed architecture of the Diff-C model, consisting of a condition encoder and a noise estimation network, in Tab. 1. The primary objective of Diff-C is to estimate the intermediate noise ϵ_θ for iterative denoising. In this architecture, $\oplus$ and $\odot$ represent concatenation and Hadamard product operations, respectively, while ϱ denotes the Softplus activation function. Fully connected layers are denoted as f, with their corresponding outputs represented as h, differentiated using subscripts for clarity.

3. Detailed Discussion on Uncertainty

To validate the superior confidence of MExD in its predictions, we employ a statistical testing approach to estimate uncertainty, inspired by [3]. Hypothesis testing offers an advantage over information-theoretic metrics in that the resulting p-value provides a more interpretable measure of the risk associated with rejecting the null hypothesis. Specifically, to quantify MExD's confidence in its predictions, we examine whether the difference between the empirical distributions of the two most probable classes across multiple posterior samples is statistically significant.

Given a *bag*, we perform 100 iterations to generate a set of posterior samples of predictive probabilities, applying a paired two-sample t-test. The null hypothesis assumes that the population means of the two groups (corresponding to the two highest probability classes) are equal, i.e., $\mu_1=\mu_2$, while the alternative hypothesis posits otherwise.

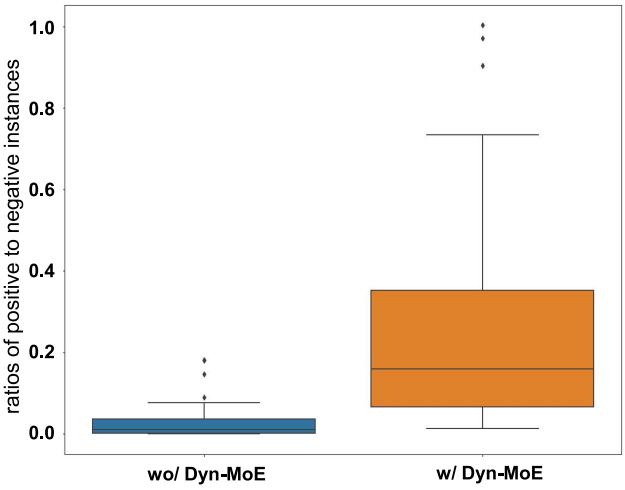

Figure 1. **Positive vs. Negative wo/w Dyn-MoE:** We present a comparative distribution of the ratios of *positive* to *negative* instances (y-axis) across all *positive* slides in the Camelyon16 [1] dataset, with and without the application of Dyn-MoE. These results demonstrate that the Dyn-MoE in MExD effectively increases the percentage of *positive* instances, highlighting its capability to address instance imbalance.

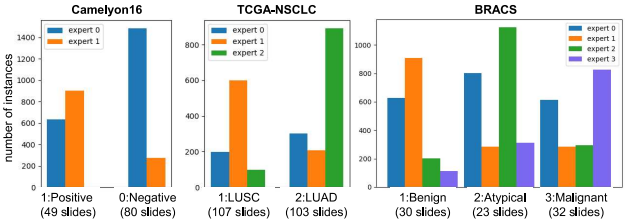

Figure 2. **Expert Load Analysis:** For slides of each subtype, we calculated the average number of instances (y-axis) assigned to each expert. The results highlight a clear correlation between subtypes and specialized experts, showcasing the tailored expertise of each expert.

In essence, a prediction rejected by the t-test indicates that the difference between the two highest probability classes is statistically significant, signifying high certainty in the prediction; conversely, failure to reject the null hypothesis suggests uncertainty.

Using the t-test results with specific significance levels (α), we can straightforwardly compute the Patch Accuracy vs Patch Uncertainty (PAvPU) [6], defined as follows:

$$\text{PAvPU} = \frac{n_{ac} + n_{iu}}{n_{ac} + n_{au} + n_{ic} + n_{iu}}, \tag{23}$$

where n_{ac}, n_{au} represent the number of predictions that are both correct and certain, and correct but uncertain, respectively. Similarly, n_{ic} and n_{iu} denote the number of predictions that are incorrect but certain, and incorrect and uncertain, respectively.

As shown in Tab. 2, we compare MExD with ACMIL [11], IBMIL [4], TransMIL [8], and MambaMIL [10]. Both ACMIL and IBMIL experience higher uncertainty in correct predictions due to their tailored instance selection processes, leading to lower PAvPU scores. In contrast, TransMIL and MambaMIL exhibit confidence in their predictions, even for incorrect cases. Notably, our MExD demonstrates certainty in all correct predictions while maintaining uncertainty for some incorrect predictions.

Table 2. **Quantitative Uncertainty Assessment:** Using PAvPU (α-value = 0.05) [6], we conducted hypothesis testing with CTransPath [9] as $\boldsymbol{f}_{\mathbb{PFE}}$. Evaluations were performed on the Camelyon16 (129 slides), TCGA-NSCLC (210 slides), and BRACS (85 slides) datasets.

Method	Dataset	n_{ac}	n_{au}	n_{ic}	n_{iu}	ACC	PAvPU
ACMIL	C16	121	1	7	0	94.57	93.80
	TCGA	197	1	12	0	94.29	93.81
	BRACS	61	2	18	4	74.12	76.47
IBMIL	C16	120	2	7	0	94.57	93.02
	TCGA	194	3	9	4	93.81	94.29
	BRACS	62	1	20	2	74.12	75.29
TransMIL	C16	122	0	7	0	94.57	94.57
	TCGA	193	0	17	0	91.90	91.90
	BRACS	62	0	23	0	72.94	72.94
MambaMIL	C16	123	0	6	0	95.35	95.35
	TCGA	199	0	11	0	94.76	94.76
	BRACS	61	0	24	0	71.76	71.76
MExD	C16	126	0	2	1	**97.67**	**98.45**
	TCGA	203	0	5	2	**96.67**	**97.62**
	BRACS	65	0	16	4	**76.47**	**81.18**

We emphasize that uncertainty metrics serve as an indicator of whether a model's prediction for each *bag* can be trusted. For incorrect and uncertain predictions, the *bag* can be referred to clinicians for further evaluation, mitigating the risk of cancerous misjudgments. This approach aligns well with the goals of human-machine collaboration in clinical settings [5].

A critical prerequisite for conducting the paired two-sample t-test is the normality assumption. To verify this, we examine Q-Q plots for the distributions of the differences between the two highest probability classes within each *bag*. As shown in the example plots in Fig. 3, all points align closely with the 45-degree line, confirming that the differences are approximately normally distributed.

4. Additional Qualitative Visualization

We provide additional visualizations of patch-wise router scores for *positive bags* to highlight the instance sparsification capability of Dyn-MoE. As illustrated in Fig. 4, the first column displays all slides, with cancerous regions outlined in green. It is evident that *positive* expert 1 in MExD effectively identifies true *positive* instances, while expert 0 predominantly focuses on true *negative* instances. This separation facilitates the partitioning of instances based on dis-

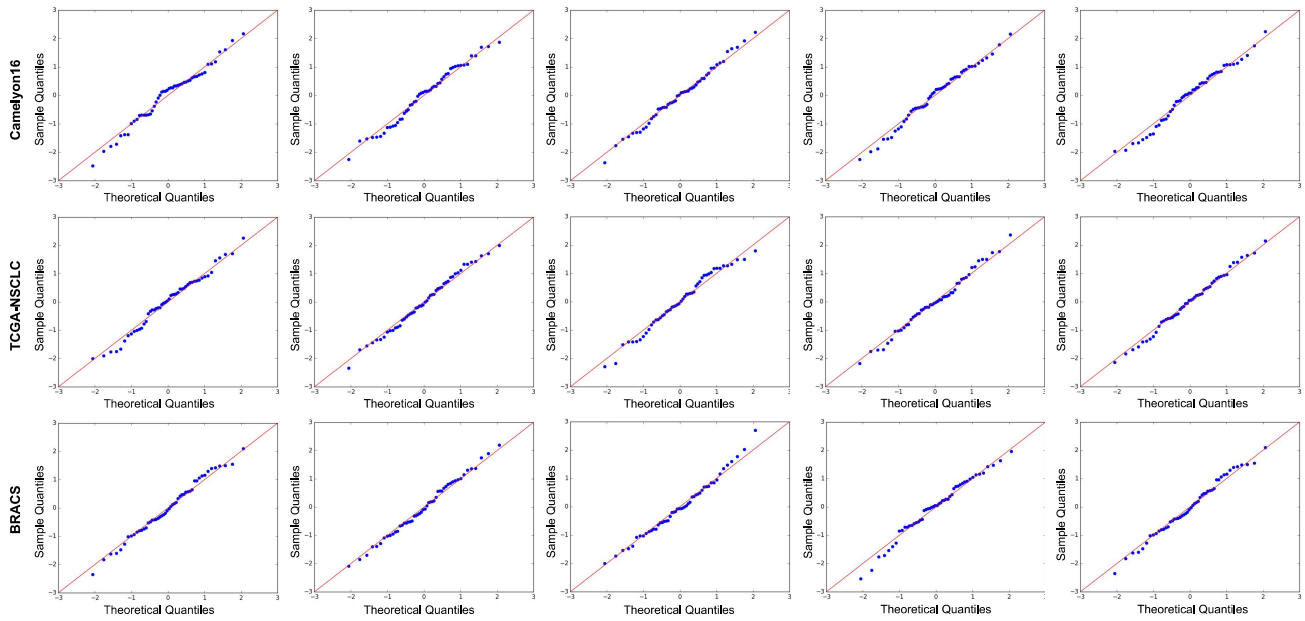

Figure 3. **Normality Assumption Assessment:** Q-Q plots illustrating the probability differences between the top two predicted classes within a *bag*, with five *bags* from each benchmark selected for visualization.

tinct properties.

Each expert then selects the most representative instances by applying a 'top-k' filtering mechanism to the router scores, using different thresholds ($\alpha_0 = 0.5$ for *positive* instances and $\alpha_1 = 0.25$ for *negative* instances). This process eliminates redundant instances and addresses the inherent imbalance between *positive* and *negative* instances. For particularly challenging slides, as shown in Fig. 4, where *positive* instances are vastly outnumbered by *negative* ones, MExD demonstrates the ability to generate accurate predictions by leveraging Dyn-MoE to sparsify the original instance sets.

To provide further insight, Fig. 1 compares the distribution of *positive* to *negative* instance ratios across all 49 *positive* slides in Camelyon16 [1]. The results indicate an increased proportion of *positive* instances in the sparse instance set compared to *negative* ones, highlighting the effectiveness of Dyn-MoE in mitigating instance imbalance.

5. Expert Load Analysis

In this section, we thoroughly examine the workload distribution across distinct experts in MExD across all benchmarks. As depicted in Fig. 2, each subtype is primarily handled by a specific expert, which proficiently selects strongly relevant instances. This specialization underscores the model's capacity to adapt dynamically without over-relying on any single expert. Instead, MExD ensures that each expert is specialized in processing its respective instances, enabling efficient handling of diverse slides through dynamic scheduling.

To validate this, we calculated the average number of instances allocated to each expert for slides belonging to a specific subtype. The details of the subtypes and corresponding datasets are as follows:

- **Camelyon16** [1]: The test split of the Camelyon16 dataset includes 49 *positive* slides and 80 *negative* slides.
- **TCGA-NSCLC**: The test split of the TCGA-NSCLC dataset comprises 107 Lung Squamous Cell Carcinoma (LUSC) slides and 103 Lung Adenocarcinoma (LUAD) slides.
- **BRACS** [2]: The test split of the BRACS dataset includes 30 Benign slides, 23 Atypical slides, and 32 Malignant slides.

These results demonstrate MExD's ability to efficiently allocate instances to experts based on their relevance to specific subtypes, ensuring balanced workload distribution and optimal performance across heterogeneous datasets.

References

[1] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 3, 4

[2] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume

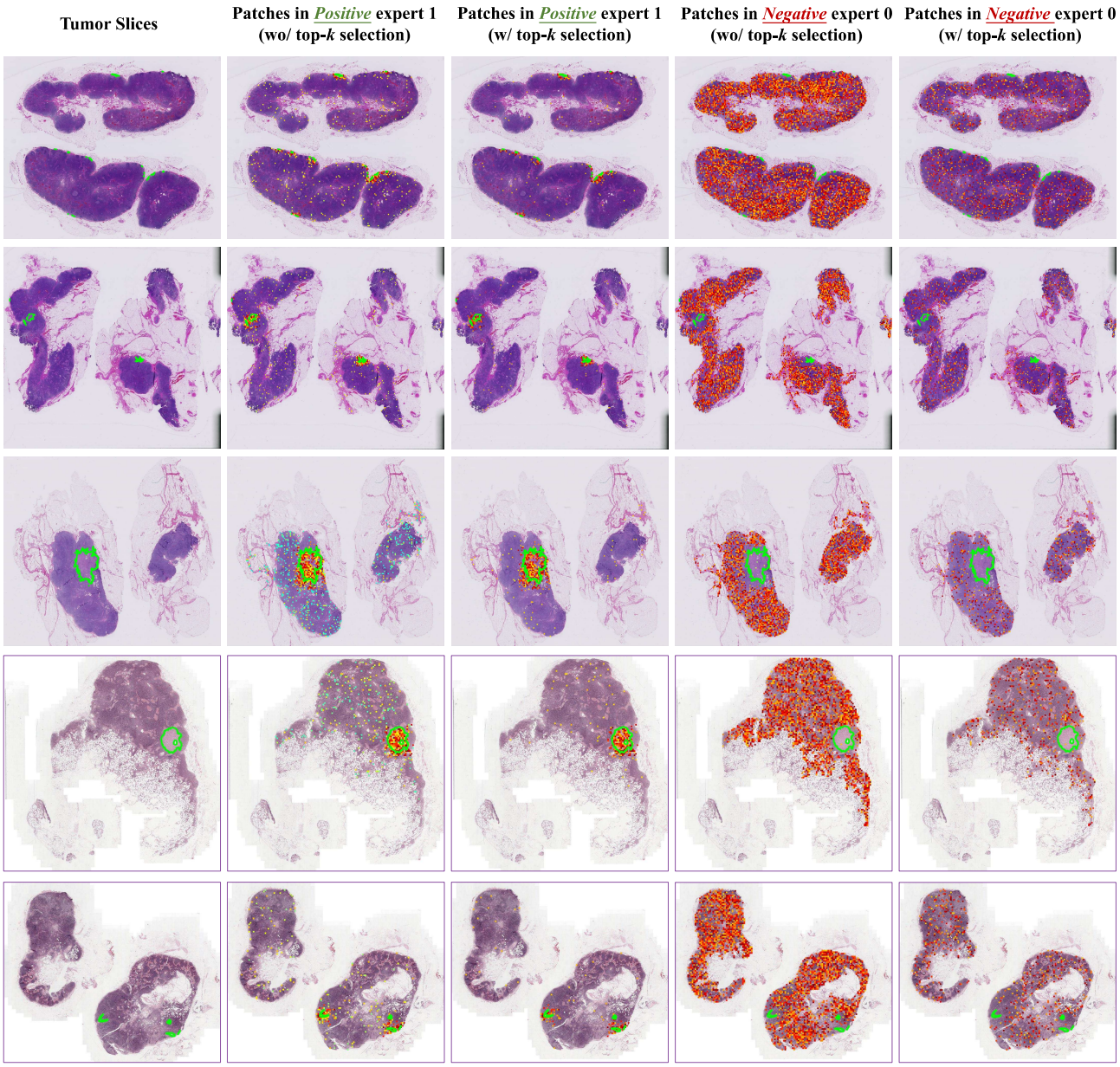

Figure 4. **Distribution Visualization** of patch-wise router scores for *positive* instances, where each score corresponds to a selected patch. In column 2, expert 1 effectively identifies *positive* patches and further refines them by retaining the most representative ones (column 3) through score-based selection. Concurrently, expert 0 focuses on refining *negative* instances. Green edges denote cancerous regions.

Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022:baac093, 2022. 4

[3] Xinjie Fan, Shujian Zhang, Korawat Tanwisuth, Xiaoning Qian, and Mingyuan Zhou. Contextual dropout: An efficient sample-dependent dropout module. *arXiv preprint arXiv:2103.04181*, 2021. 2

[4] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *CVPR*, 2023. 3

[5] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. In *NeurIPS*, 2018. 3

[6] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 3

[7] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, con-

trollable and high-fidelity generation from low-dimensional latents. *Transactions on Machine Learning Research*, 2022. 1

[8] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *NeurIPS*, 2021. 3

[9] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, 2020. 3

[10] Shu Yang, Yihui Wang, and Hao Chen. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *MICCAI*, 2024. 3

[11] Yunlong Zhang, Honglin Li, Yuxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. *arXiv preprint arXiv:2311.07125*, 2023. 3