MEC: A Multi-Expert Consultation Framework for Synergizing Frozen Foundation Models in Whole Slide Image Analysis

Zongyi Chen[1], Yu Liang[2], Weiping Lin[3], Rui Chen[4], Baptiste Magnier[5,6], Xi Wang[7], and Liansheng Wang[3,1,8(✉)]

[1] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
zongyichen@stu.xmu.edu.cn
[2] Institute of Artificial Intelligence, Xiamen University, Xiamen, China
36920251153129@stu.xmu.edu.cn
[3] Department of Computer Science, School of Informatics, Xiamen University, Xiamen, China
wplin@stu.xmu.edu.cn, lswang@xmu.edu.cn
[4] Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
drchenrui@foxmail.com
[5] EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France
baptiste.magnier@mines-ales.fr
[6] Service de Médecine Nucléaire, Centre Hospitalier Universitaire de Nîmes, University of Montpellier, Nîmes, France
[7] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China
vancywangxi@ust.hk
[8] Chongqing Jinfeng Laboratory, Chongqing, China

Abstract. The integration of Pathology Foundation Models (PFMs) with Multiple Instance Learning (MIL) has emerged as a prevailing paradigm for Whole Slide Image (WSI) analysis. However, due to diverse pre-training architectures and data, PFMs exhibit intrinsic heterogeneity and inconsistent performance across tasks. This limitation mirrors the challenge of clinical inter-observer variability, where relying on a single pathologist introduces the risk of subjective bias. To address this, we propose MEC, a Multi-Expert Consultation framework compatible with arbitrary weight-based MIL aggregators, designed to simulate the clinical consultation workflow. Specifically, MEC leverages diverse frozen PFMs as digital experts to independently localize task-specific discriminative regions within heterogeneous feature spaces. To broaden the overall scope of collective diagnostic evidence, we introduce a Stochastic Evidence Masking strategy that suppresses top-ranking features, thereby surfacing a wider variety of underlying cues. Finally, MEC synthesizes a robust, multi-view diagnostic decision by aggregating the collective consultation consensus with individual expert opinions. Extensive experiments demonstrate that MEC outperforms individual PFM baselines and alternative ensemble strategies, offering a practical framework to harness

collective diagnostic potential without requiring additional large-scale pre-training. Code is available at https://github.com/JoeyFe/MEC.

Keywords: Whole slide image · Pathology Foundation Models · Multiple Instance Learning.

1 Introduction

Computational pathology has revolutionized diagnostic workflows by facilitating automated histological sub-typing and prognostic analysis on gigapixel Whole Slide Images (WSIs) [6,21]. To mitigate the computational burden imposed by the massive resolution of WSIs, a prevailing paradigm has emerged that employs frozen Pathology Foundation Models (PFMs) for feature extraction, followed by Multiple Instance Learning (MIL) aggregation [25]. This strategy leverages the robust representation capabilities of PFMs to achieve superior performance in downstream tasks [2,15,17].

However, due to diverse pre-training data and architectures, PFMs exhibit intrinsic heterogeneity and distinct inductive biases [2]. Recent comprehensive benchmarks show that no single model is consistently superior [2,15,17], mirroring clinical inter-observer variability [5,18] where single-pathologist reliance risks subjective bias. Leveraging collaborative consensus among heterogeneous experts should yield superior performance. Existing solutions generally fall into three categories (Figure 1). First, model selection [8,26] heuristically searches for the optimal PFM-MIL combination; however, this process is computationally exhaustive and discards complementary expertise inherent in sub-optimal models. Second, feature alignment [9,13,14,27] constructs a unified feature space from multiple experts (e.g., knowledge distillation); however, these task-agnostic methods often overlook task-specific nuances and require extensive post-training resources. Third, naive ensembles [2,15] combine predictions via voting or averaging, overlooking deep feature interactions and lacking robustness as noise from weaker experts frequently dilutes reliable insights.

Fundamentally, existing ensemble approaches function by directly fusing latent features or final predictions in a black-box manner. Such a process frequently overlooks that the localization of key diagnostic evidence represents a crucial component of both clinical workflows [10,16] and MIL paradigms [7,12]. To systematically capture the diverse pathological cues essential for multifaceted clinical diagnosis, we propose MEC, a Multi-Expert Consultation framework that emulates the workflow of evidence-based joint diagnosis. In contrast to conventional ensembles, MEC employs PFMs as digital experts, enabling the independent localization of task-specific regions to capture multi-view evidence. Furthermore, we introduce a Stochastic Evidence Masking (SEM) strategy that encourages the discovery of a broader range of complementary patterns. Finally, MEC dynamically aggregates the opinions of individual experts to produce a robust diagnostic consensus.

Our contributions are summarized as follows: (1) We propose MEC, a generic WSI analysis framework compatible with arbitrary weight-based MIL aggrega-

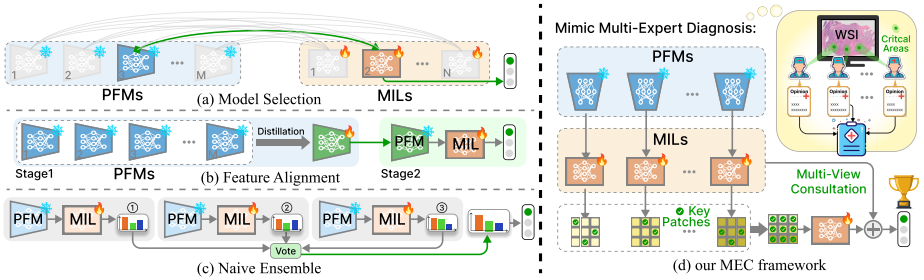

Fig. 1: Comparison of (a) Model Selection, (b) Feature Alignment, (c) Naive Ensembles, and (d) our MEC framework. MEC mimics clinical joint diagnosis, dynamically consulting diverse experts for an evidence-driven consensus.

tors. By leveraging frozen foundation models, it efficiently adapts to downstream tasks without requiring additional large-scale training. (2) We reveal the inherent complementarity of distinct PFMs in Region of Interest (ROI) mining and accordingly introduce a multi-view key patch mining strategy that fuses diverse evidence to construct robust task-specific representations. (3) Extensive benchmarking across multiple datasets demonstrates the consistent superiority of MEC over individual PFM baselines and ensemble strategies, validating it as a practical, general-purpose framework for harnessing collective diagnostic potential.

2 Method

The proposed Multi-Expert Consultation framework structurally simulates the clinical consensus process by synergizing heterogeneous PFMs to achieve robust and multi-view WSI analysis. The overall architecture is illustrated in Figure 2.

2.1 Feature Extraction with Heterogeneous PFMs

Following standard WSI preprocessing protocols [12], the tissue area of a WSI is tessellated into N non-overlapping patches $\{p_i\}_{i=1}^{N}$. To exploit complementary representations, we employ a set of M heterogeneous and frozen PFMs, denoted by $\{f_m\}_{m=1}^{M}$. Each expert model f_m projects the patches into a distinct feature space $\mathbb{R}^{D_m}$. Specifically, $\mathbf{H}^{(m)} = [\mathbf{h}_1^{(m)}, \ldots, \mathbf{h}_N^{(m)}]^\top \in \mathbb{R}^{N \times D_m}$ represents the bag of features derived from the m-th expert, where $\mathbf{h}_i^{(m)} \in \mathbb{R}^{D_m}$ denotes the embedding of the i-th patch.

2.2 Task-Specific Evidence Localization

We employ a multi-branch architecture, where each frozen PFM is paired with a learnable MIL backbone. For the m-th branch, the feature bag $\mathbf{H}^{(m)} \in \mathbb{R}^{N \times D_m}$ is processed by a dedicated MIL network. Our framework accommodates diverse

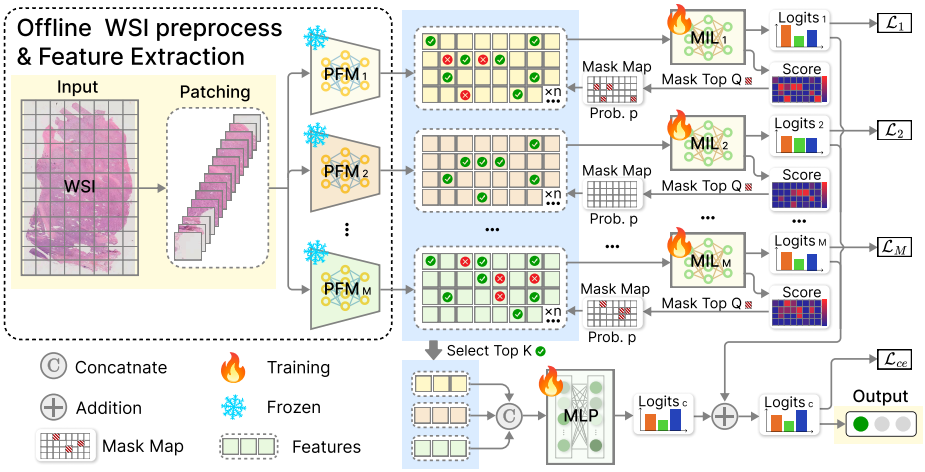

Fig. 2: Overview of the MEC framework utilizing frozen foundation model features as input for task-specific adaptation.

MIL architectures that produce instance-level importance weights. Let $\Phi_m(\cdot;\theta_m)$ denote the instance scoring network (e.g., the widely adopted attention network) of the m-th expert. This network maps input embeddings to importance scores $\mathbf{A}^{(m)} \in \mathbb{R}^{1\times N}$, quantifying the diagnostic contribution of each patch:

$$\mathbf{A}^{(m)} = \Phi_m(\mathbf{H}^{(m)};\theta_m). \tag{1}$$

To drive the learning of task-aware representations, we impose independent supervision on each expert branch. Specifically, within the m-th branch, weighted patch-level features are aggregated into a slide-level representation $\mathbf{z}_{slide}^{(m)}$ via a global function $\mathcal{G}_m(\cdot)$, which accommodates diverse MIL architectures:

$$\mathbf{z}_{slide}^{(m)} = \mathcal{G}_m(\mathbf{H}^{(m)}, \mathbf{A}^{(m)};\theta_m). \tag{2}$$

Subsequently, the embedding $\mathbf{z}_{slide}^{(m)}$ is fed into a classifier $g_m(\cdot)$ to generate logits $\hat{\mathbf{y}}^{(m)}$. We enforce discriminative learning by applying a Cross-Entropy loss $\mathcal{L}_{CE}$ to each expert:

$$\mathcal{L}_{experts} = \frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_{CE}(\hat{\mathbf{y}}^{(m)}, y), \tag{3}$$

where y denotes the ground-truth label. This explicit supervision aligns feature aggregation with the ground truth, yielding more reliable importance scores $\mathbf{A}^{(m)}$ for subsequent mining.

2.3 Stochastic Evidence Masking

We introduce a stochastic masking strategy inspired by regularization techniques [28,29]. While prior works employ this mechanism to prevent overfit-

ting in single-view MIL, we repurpose it to expand the collective pool of diagnostic evidence. Specifically, in the training phase, we generate a binary mask $\mathbf{M}^{(m)} \in \{0,1\}^N$ to suppress the instances with the top-Q scores with a probability p. The masked importance scores $\tilde{\mathbf{A}}^{(m)}$ are then obtained via element-wise multiplication:

$$\tilde{\mathbf{A}}^{(m)} = \mathbf{A}^{(m)} \odot \mathbf{M}^{(m)}. \tag{4}$$

By preventing the model from collapsing onto a narrow set of dominant features, this perturbation ensures a richer basis for the subsequent multi-view integration.

2.4 Synthesizing Multi-View Predictions

Following the evidence mining phase, we aggregate multi-view diagnostic cues to establish a holistic consensus. Importance score distributions in MIL are inherently sparse, with a small subset of highly discriminative patches dominating the slide-level prediction [22,28]. Motivated by this observation, we focus on capturing these critical instances. Specifically, for the m-th expert branch, we utilize the masked importance scores $\tilde{\mathbf{A}}^{(m)}$ (derived in Eq. 4) to retrieve the top-K most critical instances. Let $\mathcal{I}^{(m)}$ denote the indices corresponding to the K highest values in $\tilde{\mathbf{A}}^{(m)}$, sorted in descending order. To fully leverage the rich semantic representations inherent in the foundation models, we extract the raw feature embeddings from the original feature bag $\mathbf{H}^{(m)}$ at these indices to construct a sequence of critical evidence $\mathbf{S}^{(m)} = [\mathbf{h}_i^{(m)}]_{i \in \mathcal{I}^{(m)}} \in \mathbb{R}^{K \times D_m}$. These selected instance features $\mathbf{S}^{(m)}$ are subsequently flattened into a vector $\mathbf{v}^{(m)} \in \mathbb{R}^{K \cdot D_m}$. To synthesize a unified representation, we concatenate the flattened vectors from all M experts:

$$\mathbf{V}_{fused} = \text{Concat}\left([\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(M)}]\right). \tag{5}$$

The fused vector $\mathbf{V}_{fused}$, which encapsulates complementary evidence derived from diverse inductive biases, is processed by a Multi-Layer Perceptron (MLP) to obtain a high-level consensus representation. We formulate this feature transformation and the subsequent classification as:

$$\mathbf{z}_{fused} = \mathcal{F}_{mlp}(\mathbf{V}_{fused}; \Theta_{mlp}), \quad \hat{\mathbf{y}}_{consensus} = \mathcal{G}_{cls}(\mathbf{z}_{fused}), \tag{6}$$

where $\mathcal{F}_{mlp}(\cdot)$ denotes the non-linear mapping parameterized by Θ_{mlp}, and $\mathcal{G}_{cls}(\cdot)$ represents the final classification head projecting the refined features $\mathbf{z}_{fused}$ to the target class logits. Inspired by the bias-variance decomposition principles underlying ensemble learning [4], we employ a residual-style aggregation strategy in logit space to derive the final slide-level prediction $\hat{\mathbf{y}}_{final}$. This design fuses the high-capacity collaborative consensus $\hat{\mathbf{y}}_{consensus}$ with the stable ensemble of individual expert predictions, yielding a more robust final decision.

$$\hat{\mathbf{y}}_{final} = \hat{\mathbf{y}}_{consensus} + \frac{1}{M} \sum_{m=1}^{M} \hat{\mathbf{y}}^{(m)}. \tag{7}$$

The total objective function $\mathcal{L}_{total}$ is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{CE}(\hat{\mathbf{y}}_{final}, y) + \mathcal{L}_{experts}. \tag{8}$$

Table 1: Performance evaluation of different feature extractors combined with various MIL models across diverse pathological tasks. Best results are highlighted in **bold**, and the best single-model results are underlined.

Dataset		Camelyon17		TCGA-GBMLGG		TCGA-NSCLC		STS	
	Method	Acc	F1-score	Acc	F1-score	Acc	F1-score	Acc	F1-score
ABMIL	UNI	79.17±3.88	50.45±3.30	83.25±2.29	80.80±2.61	91.95±1.60	91.91±1.63	69.12±7.73	58.14±5.11
	CONCHv1.5	83.53±1.53	57.74±4.07	83.69±2.60	81.21±2.85	92.71±1.92	92.68±1.95	72.57±7.02	61.78±3.68
	Virchow2	83.79±2.29	59.55±4.63	83.31±2.94	81.01±4.10	92.97±1.30	92.95±1.31	73.30±7.11	64.95±3.38
CLAM	UNI	83.37±2.96	56.43±1.90	82.66±1.60	80.02±1.74	92.21±1.51	92.19±1.52	69.68±8.62	56.08±6.01
	CONCHv1.5	83.36±1.15	57.51±2.12	81.37±1.38	78.64±1.64	92.04±1.73	92.01±1.76	69.00±10.95	56.07±7.59
	Virchow2	84.79±2.30	59.57±1.67	83.37±1.86	81.06±2.09	93.36±1.45	93.33±1.47	73.39±7.38	63.52±6.25
TransMIL	UNI	69.50±3.44	38.96±5.22	80.86±2.76	78.18±3.05	89.97±3.21	89.92±3.26	70.52±6.86	60.87±3.70
	CONCHv1.5	82.99±3.34	60.67±3.96	85.12±2.69	82.51±2.14	90.42±1.44	90.39±1.45	70.01±7.63	59.91±6.32
	Virchow2	71.13±5.52	44.51±7.10	83.21±4.27	81.06±3.93	91.38±2.28	91.36±2.31	71.07±7.85	61.72±5.78
RRTMIL	UNI	78.13±5.95	51.68±7.38	83.90±3.12	81.86±3.43	92.03±1.47	92.02±1.49	71.28±7.12	60.90±4.99
	CONCHv1.5	81.95±4.92	60.69±3.18	85.05±2.81	82.78±2.08	92.04±2.34	92.02±2.37	70.73±9.37	60.54±5.76
	Virchow2	82.22±6.17	58.16±7.11	85.20±4.35	83.33±4.01	92.41±1.66	92.39±1.68	70.84±8.59	60.21±4.95
TDAMIL	UNI	73.84±5.91	47.82±12.24	85.21±2.01	83.33±1.57	90.89±1.65	90.85±1.68	70.84±7.71	60.29±4.62
	CONCHv1.5	83.79±5.19	59.40±4.47	84.12±2.67	81.91±2.01	91.57±2.33	91.53±2.38	71.90±6.15	61.84±7.58
	Virchow2	80.21±3.37	55.38±4.19	85.49±3.52	83.87±2.87	92.32±1.61	92.28±1.64	71.81±7.50	63.11±4.19
MEC$_{\text{ABMIL}}$		**86.96±3.10**	**68.44±2.64**	87.79±1.92	86.05±2.20	93.94±1.72	93.91±1.75	77.16±5.80	68.72±4.23
MEC$_{\text{ABMIL-G}}$		85.20±2.47	64.26±1.96	87.61±1.69	86.08±1.96	93.75±1.61	93.72±1.62	76.56±4.26	68.87±5.41
MEC$_{\text{CLAM}}$		86.31±3.83	66.90±4.07	**87.80±2.55**	**86.20±2.69**	**94.04±2.04**	**94.02±2.05**	**77.16±5.52**	**70.20±5.06**

3 Experiments

Datasets. We evaluate the proposed framework on four real-world datasets: Camelyon17 (500 WSIs, four-class lymph node metastasis detection) [1], TCGA-NSCLC (1,053 WSIs, binary lung cancer subtyping) [12], TCGA-GBMLGG (1,525 WSIs, three-class glioma molecular subtyping) [24], and STS (813 WSIs, fine-grained classification of five soft tissue sarcoma subtypes). We use a standard 5-fold cross-validation protocol, ensuring no patient overlap across training, validation, and testing sets.

Implementation Details. Following [15], we use PrePATH for WSI preprocessing, tissue regions are segmented into non-overlapping 512 × 512 patches at 20× magnification. For feature extraction, we employ three frozen pathology foundation models: UNI [3], Virchow2 [30], and CONCHv1.5 [11]. For the consensus branch, the concatenated evidence vector $\mathbf{V}_{fused}$ is fed into an MLP classifier with two hidden layers of dimensions 512 and 128, where each hidden layer is followed by a ReLU activation and dropout with a rate of 0.3. To ensure fair comparisons, all methods are optimized using the Adam optimizer coupled with a cosine annealing learning rate schedule.

Comparison with Single-PFM Baselines. We benchmark 15 combinations, pairing three PFMs (UNI, CONCHv1.5, Virchow2) with five MIL architectures

Table 2: Performance comparison of different ensemble methods across diverse pathological tasks. Best in **bold**, second best is underlined.

Dataset	Camelyon17		TCGA-GBMLGG		TCGA-NSCLC		STS	
Method	Acc	F1-score	Acc	F1-score	Acc	F1-score	Acc	F1-score
Mean	$85.00_{\pm 2.44}$	$63.91_{\pm 7.05}$	$\underline{86.66}_{\pm 2.36}$	$84.75_{\pm 2.81}$	$92.98_{\pm 1.38}$	$92.94_{\pm 1.39}$	$\underline{75.32}_{\pm 5.35}$	$65.62_{\pm 3.62}$
Concat	$83.19_{\pm 3.64}$	$60.06_{\pm 4.18}$	$86.44_{\pm 2.20}$	$\underline{84.76}_{\pm 2.26}$	$92.98_{\pm 1.55}$	$92.94_{\pm 1.58}$	$75.18_{\pm 5.49}$	$\underline{65.70}_{\pm 3.33}$
Logit	$\underline{85.78}_{\pm 3.15}$	$\underline{64.59}_{\pm 2.51}$	$84.29_{\pm 2.72}$	$81.84_{\pm 3.40}$	$93.17_{\pm 1.55}$	$93.14_{\pm 1.57}$	$74.37_{\pm 6.50}$	$64.99_{\pm 5.41}$
Vote	$85.42_{\pm 2.53}$	$61.77_{\pm 2.85}$	$85.05_{\pm 3.08}$	$82.79_{\pm 1.26}$	$\underline{93.26}_{\pm 1.43}$	$\underline{93.24}_{\pm 1.14}$	$74.45_{\pm 6.59}$	$64.84_{\pm 2.61}$
COBRA	$75.91_{\pm 3.35}$	$43.16_{\pm 2.83}$	$85.05_{\pm 1.82}$	$83.31_{\pm 1.21}$	$92.33_{\pm 1.78}$	$92.30_{\pm 1.81}$	$72.93_{\pm 8.22}$	$65.61_{\pm 7.49}$
GPFM	$82.78_{\pm 2.32}$	$60.36_{\pm 3.52}$	$86.26_{\pm 2.91}$	$84.29_{\pm 3.62}$	$92.97_{\pm 2.03}$	$92.93_{\pm 2.03}$	$74.21_{\pm 6.40}$	$65.37_{\pm 6.40}$
MEC	$\mathbf{86.96}_{\pm 3.10}$	$\mathbf{68.44}_{\pm 2.64}$	$\mathbf{87.79}_{\pm 1.92}$	$\mathbf{86.05}_{\pm 2.20}$	$\mathbf{93.94}_{\pm 1.72}$	$\mathbf{93.91}_{\pm 1.75}$	$\mathbf{77.16}_{\pm 5.80}$	$\mathbf{68.72}_{\pm 4.23}$

* We use official pre-trained weights to extract features for COBRA and GPFM. COBRA follows the recommended linear probing method, whereas ABMIL is used as the backbone for all other methods.

(ABMIL [7], CLAM [12], TransMIL [20], RRTMIL [23], TDAMIL [19]). As presented in Table 1, our MEC framework consistently outperforms all single-PFM baselines. These results support that integrating complementary inductive biases can alleviate the limitations of any individual PFM. Besides, consistent with prior observations [3,14,26], performance varies only marginally with the choice of MIL architecture when expressive PFMs are used. This observation suggests that rich features enable standard attention mechanisms to localize evidence effectively, thereby justifying our use of simple attention scores for patch mining. Finally, the consistent improvements observed with ABMIL-GATE and CLAM demonstrate that our framework remains MIL-agnostic, effectively supporting diverse aggregation networks without altering the consensus mechanism.

Comparison with Ensemble Methods. We compare MEC with feature-level fusion (Mean, Concat), prediction-level fusion (Logit, Vote), and feature alignment models (GPFM [14], COBRA [9]). As shown in Table 2, MEC consistently outperforms all baselines. Its superiority over static ensembles validates our dynamic consensus mechanism in effectively filtering noise and amplifying discriminative cues across diverse expert representations, whereas surpassing COBRA/GPFM proves the efficacy of task-driven probing without heavy pretraining. **In terms of efficiency**, multi-PFM feature extraction increases offline preprocessing time (49.8 s/WSI on a single GPU vs. 28.0 s for GPFM), but this cost is highly parallelizable (30.3 s/WSI with multi-GPU). Crucially, MEC's MIL inference requires merely 2.02 ms. This renders the accuracy-compute trade-off highly justified for offline clinical WSI workflows, where maximizing diagnostic precision on complex tasks strictly outweighs the need for absolute latency.

Ablation Study. To validate the architectural design of MEC, we compare the full framework against three variants: (1) *w/o consen.*, which removes the consensus branch and relies solely on aggregated expert logits; (2) *w/o experts*, which utilizes only the consensus branch prediction; and (3) *w/o SEM*, which removes the stochastic evidence masking mechanism. As presented in Figs. 3a and 3b, the full MEC framework consistently outperforms all ablated variants

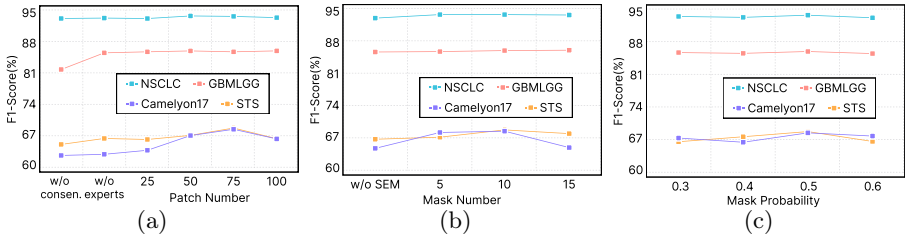

Fig. 3: Ablation study on key components. (a) Impact of instance selection number K. (b) Impact of mask number. (c) Impact of mask probability.

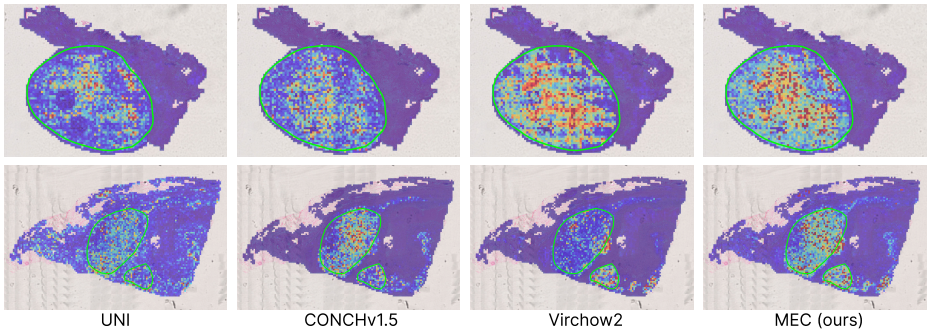

Fig. 4: Visualization of WSI sample heatmaps on the STS dataset using an ABMIL backbone with single foundation models versus MEC. pathologist-annotated Regions of Interest (ROIs) are circled in green.

across all datasets. This confirms that the synergy between expert guidance and consensus recalibration is critical for robust diagnosis.

We further perform ablation studies on instance selection (K). As presented in Fig. 3a, TCGA-NSCLC and TCGA-GBMLGG achieve optimal performance at $K = 50$, whereas Camelyon17 and STS favor a larger context ($K = 75$). Overall, performance is relatively stable across a reasonable range of K, suggesting that MEC is not overly sensitive to the exact threshold as long as sufficient contextual evidence is retained. Regarding masking intensity in SEM (Figs. 3b and 3c), results consistently favor moderate regularization. In practice, setting the mask number to 10 and the mask probability to 50% is generally sufficient for achieving near-optimal performance.

Analysis of Attention Regions. To qualitatively assess our collaborative mechanism, we visualize attention heatmaps produced by individual PFMs and by MEC. As illustrated in Fig. 4, different PFMs attend to markedly different regions, reflecting model-specific preferences. In contrast, MEC yields a more comprehensive and coherent attention distribution that covers complementary diagnostically relevant areas. This suggests that the consultation-based consensus

can synthesize evidence across experts, reducing reliance on any single model's bias and leading to more robust slide-level decisions.

4 Conclusions

We propose MEC, a generic framework simulating clinical multi-expert consultation to address PFM heterogeneity. Treating frozen PFMs as digital experts with distinct inductive biases, MEC employs a multi-branch architecture to parallelly localize task-specific regions. Extensive experiments demonstrate that MEC consistently outperforms individual PFMs and standard ensembles. Results validate MEC as a practical, computation-efficient solution to harness the collective potential of foundation models without extra large-scale pre-training.

Acknowledgments. This work was supported by the Ministry of Science and Technology of the People's Republic of China (STI2030-Major Projects 2021 ZD0201900), the National Natural Science Foundation of China (Grant No. 62371409), and the Fujian Provincial Natural Science Foundation of China (Grant No. 2023J01005).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al.: 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 (2019)
2. Bareja, R., Carrillo-Perez, F., Zheng, Y., Pizurica, M., Nandi, T.N., Shen, J., Madduri, R., Gevaert, O.: Evaluating vision and pathology foundation models for computational pathology: A comprehensive benchmark study. medRxiv (2025)
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
4. Dietterich, T.G.: Ensemble methods in machine learning. In: Multiple Classifier Systems. pp. 1–15. Springer Berlin Heidelberg (2000)
5. Elmore, J.G., Longton, G.M., Carney, P.A., Geller, B.M., Onega, T., Tosteson, A.N.A., Nelson, H.D., Pepe, M.S., Allison, K.H., Schnitt, S.J., O'Malley, F.P., Weaver, D.L.: Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA **313**(11), 1122–1132 (2015)
6. Hosseini, M.S., Bejnordi, B.E., Trinh, V.Q.H., Chan, L., Hasan, D., Li, X., Yang, S., Kim, T., Zhang, H., Wu, T., Chinniah, K., Maghsoudlou, S., Zhang, R., Zhu, J., Khaki, S., Buin, A., Chaji, F., Salehi, A., Nguyen, B.N., Samaras, D., Plataniotis, K.N.: Computational pathology: A survey review and the way forward. Journal of Pathology Informatics **15**, 100357 (2024)
7. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: ICML. vol. 80, pp. 2127–2136. PMLR (2018)

8. Lee, J., Lim, J., Byeon, K., Kwak, J.T.: Benchmarking pathology foundation models: Adaptation strategies and scenarios. Computers in Biology and Medicine **190**, 110031 (2025)
9. Lenz, T., Neidlinger, P., Ligero, M., Wölflein, G., van Treeck, M., Kather, J.N.: Unsupervised foundation model-agnostic slide-level representation learning. In: CVPR. pp. 30807–30817 (2025)
10. Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P., Ellison, D.W.: The 2016 world health organization classification of tumors of the central nervous system: A summary. Acta Neuropathologica **131**(6), 803–820 (2016)
11. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024)
12. Lu, M.Y., Williamson, D.F.K., 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)
13. Luo, X., Wang, X., Eweje, F., et al.: Ensemble learning of foundation models for precision oncology. arXiv preprint 2508.16085 (2025)
14. Ma, J., Guo, Z., Zhou, F., et al.: A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. Nature Biomedical Engineering (2025)
15. Ma, J., Xu, Y., Zhou, F., et al.: Pathbench: A comprehensive comparison benchmark for pathology foundation models towards precision oncology. arXiv preprint arXiv:2505.20202 (2025)
16. Mercan, E., Aksoy, S., Shapiro, L.G., Weaver, D.L., Brunyé, T.T., Elmore, J.G.: Localization of diagnostically relevant regions of interest in whole slide images: A comparative study. Journal of Digital Imaging **29**(4), 496–506 (2016)
17. Neidlinger, P., El Nahhas, O.S.M., Muti, H.S., Lenz, T., Hoffmeister, M., Brenner, H., van Treeck, M., Langer, R., Dislich, B., Behrens, H.M., Röcken, C., Foersch, S., Truhn, D., Marra, A., Saldanha, O.L., Kather, J.N.: Benchmarking foundation models as feature extractors for weakly supervised computational pathology. Nature Biomedical Engineering (2025)
18. Niazi, M.K.K., Parwani, A.V., Gurcan, M.N.: Digital pathology and artificial intelligence. The Lancet Oncology **20**(5), e253–e261 (2019)
19. Reisenbüchler, D., Deng, R., Matek, C., Feuerhake, F., Merhof, D.: Top-Down Attention-based Multiple Instance Learning for Whole Slide Image Analysis . In: MICCAI. vol. LNCS 15960, pp. 651 – 660 (2025)
20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., zhang, y.: TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In: Advances in Neural Information Processing Systems. vol. 34, pp. 2136–2147 (2021)
21. Song, A.H., Jaume, G., Williamson, D.F.K., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023)
22. 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)
23. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: CVPR. pp. 11343–11352 (2024)

24. Wang, W., Zhao, Y., Teng, L., et al.: Neuropathologist-level integrated classification of adult-type diffuse gliomas using deep learning from whole-slide pathological images. Nature Communications **14**(1), 6359 (2023)
25. Xiong, C., Chen, H., Sung, J.J.Y.: A survey of pathology foundation model: Progress and future directions. arXiv preprint arXiv:2504.04045 (2025)
26. Xu, H., Wang, M., Shi, D., Qin, H., Zhang, Y., Liu, Z., Madabhushi, A., Gao, P., Cong, F., Lu, C.: When multiple instance learning meets foundation models: Advancing histological whole slide image analysis. Medical Image Analysis **101**, 103456 (2025)
27. Yu, Z., Zhang, S., Qiao, N., Zhao, Y., Yu, L., Peng, T., Zhang, X.Y.: FM2: Fusing multiple foundation models for pathology image analysis via disentangled consensus-divergence representation. Information Fusion **127**, 103840 (2026)
28. Zhang, Y., Li, H., Sun, Y., Zheng, S., Zhu, C., Yang, L.: Attention-challenging multiple instance learning for whole slide image classification. In: ECCV. pp. 125–143 (2024)
29. Zhu, W., Qiu, P., Chen, X., Yang, Z., Sotiras, A., Razi, A., Wang, Y.: How effective can dropout be in multiple instance learning ? In: ICML. vol. 267, pp. 80090–80106. PMLR (2025)
30. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Fuchs, T., Fusi, N., Liu, S., Severson, K.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)