MOC: Meta-Optimized Classifier for Few-Shot Whole Slide Image Classification

Tianqi Xiang[1], Yi Li[1], Qixiang Zhang[1], and Xiaomeng Li[1,⋆]

Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
eexmli@ust.hk

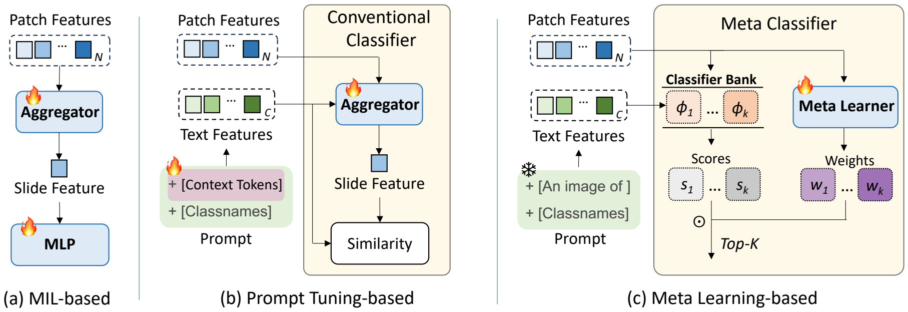
Abstract. Recent advances in histopathology vision-language foundation models (VLFMs) have shown promise in addressing data scarcity for whole slide image (WSI) classification via zero-shot adaptation. However, these methods remain outperformed by conventional multiple instance learning (MIL) approaches trained on large datasets, motivating recent efforts to enhance VLFM-based WSI classification through few-shot learning paradigms. While existing few-shot methods improve diagnostic accuracy with limited annotations, their reliance on conventional classifier designs introduces critical vulnerabilities to data scarcity. To address this problem, we propose a Meta-Optimized Classifier (MOC) comprising two core components: (1) a meta-learner that automatically optimizes a classifier configuration from a mixture of candidate classifiers and (2) a classifier bank housing diverse candidate classifiers to enable a holistic pathological interpretation. Extensive experiments demonstrate that MOC outperforms prior arts in multiple few-shot benchmarks. Notably, on the TCGA-NSCLC benchmark, MOC improves AUC by 10.4% over the state-of-the-art few-shot VLFM-based methods, with gains up to 26.25% under 1-shot conditions, offering a critical advancement for clinical deployments where diagnostic training data is severely limited. Code is available at https://github.com/xmed-lab/MOC.

Keywords: Few-Shot Learning · Whole Slide Image · Meta Learning

1 Introduction

Recent advances in pathology vision-language foundation models (VLFMs), such as PLIP [4], BiomedCLIP [19], and CONCH [10], have demonstrated remarkable capabilities in histopathological image interpretation through visual-language alignment. These models mitigate data scarcity challenges stemming from annotation costs, privacy constraints, and rare disease prevalence via zero-shot adaptation. However, existing VLFM-based zero-shot approaches exhibit inferior performance compared to conventional multiple instance learning (MIL) frameworks that leverage extensive annotated whole slide image (WSI) datasets. This performance discrepancy has motivated emerging research into few-shot VLFM-based

[⋆] Corresponding Author



(a) MIL-based (b) Prompt Tuning-based (c) Meta Learning-based

Fig. 1: Architectural comparison among (a) MIL-based methods, (b) prompt tuning-based methods, and (c) our proposed meta learning-based method for VLFM few-shot pathology analysis. Different from prompt-tuning-based frameworks that adopt **conventional classifiers**, our method uses the meta learner to dynamically compose an optimal **meta classifier** from the classifier bank.

WSI classifications that enhance diagnostic reliability with minimal annotated data.

Existing VLFM-based few-shot WSI classification methods (Fig. 1(b)), distinguish themselves from conventional MIL frameworks [5,12,15,6,18,20,9,17] (Fig. 1(a)) by incorporating linguistic supervision alongside visual understanding. Most VLFM-based few-shot WSI methods [13,7,14,3,2,1,8] address few-shot classification by introducing various prompt-tuning techniques. For example, TOP [13] leverages pathology prior knowledge in its prompting, FiVE [7] integrates pathological reports into prompts, PEMP [14] enhances prompts with additional image references, and FAST [2] employs supplementary visual prompts through cached samples. However, these methods mainly focus on prompt engineering at the input side while adopting the conventional classifier comprising a learnable aggregator (e.g., attention pooling) to generate global representations and a visual-linguistic similarity matching for predictions. Our experimental analysis reveals this design's vulnerability for few-shot learning, with TOP [13] exhibiting a 25% AUC drop (0.79→0.54) by decreasing the training samples from 16 to 2 (see Tab. 1). Such performance degradation is primarily due to the overfitting of parameter-intensive aggregators under data scarcity.

In light of this, we consider using non-parametric operations as an alternative to the attention-based aggregator. Notably, our baseline uses non-parametric top-K similarity matching to achieve promising performance, surpassing state-of-the-art prompt-tuning methods (CoOp [21] and TOP [13]) by at least 4.9% in AUC across various few-shot settings (see Tab. 1). This finding highlights the critical role of classifier design, i.e., cosine similarity, in few-shot scenarios. However, relying solely on top-K similarity matching, which prioritizes patches with maximal disease descriptor alignment, may yield suboptimal outcomes. Therefore, we further propose a meta-learning-based method that jointly leverages multiple classifiers with complementary diagnostic emphases, enabling a more effective identification of the most representative patches from the slide for diagnosis.

In this work, we introduce a novel Meta-Optimized Classifier (MOC) for few-shot whole slide image classification. MOC comprises two core components: (1) a meta-learner that automatically optimizes a classifier configuration from a mixture of candidate classifiers and (2) a classifier bank housing diverse candidate classifiers to enable a holistic pathological interpretation. Specifically, the classifier bank is empirically implemented with four non-parametric operations, emphasizing three key aspects: maximal similarity, categorical prominence, and minimal irrelevance. Extensive experiments demonstrate that MOC consistently outperforms prior art across multiple few-shot benchmarks. Notably, on the TCGA-NSCLC benchmark, MOC achieves a significant improvement of 10.4% in AUC over the state-of-the-art few-shot methods, with performance gains amplifying to 26.25% under 1-shot conditions. This advancement is particularly critical for clinical deployments, where diagnostic training data is severely limited.

2 Methodology

In this section, we introduce a novel few-shot WSI classification framework centered around our Meta-Optimized Classifier (MOC). As illustrated in Fig. 2, the framework begins with WSI preprocessing, followed by the application of MOC, which leverages a meta-learner to dynamically propose configurations and integrate candidate classifiers from the classifier bank. We will first elaborate on how the MOC assists few-shot WSI classification, followed by our construction recipe for the classifier bank.

The problem definition of our few-shot WSI classification task is briefly presented as follows: given a dataset $\mathcal{D} = \{X_1, X_2, ..., X_N\}$ comprising N WSIs with C distinct categories, each WSI X_i is given a label $Y_i \in \{1, 2, ..., C\}$. Each WSI X_i is then partitioned into n_i non-overlapping patches $\{x_{i,j}, j = 1, 2, ..., n_i\}$, where the label for each patch $x_{i,j}$ is unknown. "Shot" refers to the number of labeled WSIs for each category, i.e., the training set for C-Category K-Shot WSI classification contains $K \times C$ labeled WSIs. Typically, K can take values such as 1, 2, 4, or 8.

2.1 Meta-Optimized Classifier for Few-Shot WSI Classification

The core function of the MOC is to construct an optimal classifier for each input instance. This is accomplished by the collaboration of the meta-learner and the classifier bank. Candidates in the classifier banks provide pathological analysis from different aspects, while the meta-learner coordinates such observations for an optimal classifier scheme.

To be specific, we firstly *preprocess* the WSIs. We utilize the pathology vision-language foundation model as the backbone of the proposed framework, which

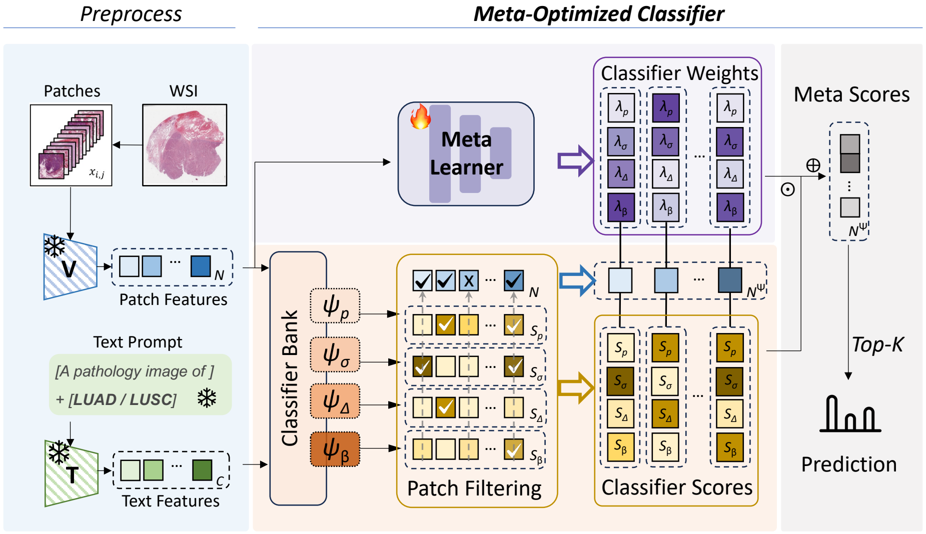


Fig. 2: The two-phase few-shot WSI classification pipeline: 1) Preprocess WSI and prompts. 2) WSI prediction with the proposed Meta-Optimized Classifier.

consists of pre-trained visual encoder $\mathcal{F}(\cdot)$ and text encoder $\mathcal{G}(\cdot)$. For each category c, the prompt t_c = template + classname$_c$, where the template in the form of `‘A pathology image of {}’` and the classname$_c$ takes this full name of the category, i.e., `‘lung adenocarcinoma’`. Then, the corresponding prompt embeddings can be computed via $w_c = \frac{\mathcal{G}(t_c)}{\|\mathcal{G}(t_c)\|}$, and the l_2-normalized WSI embeddings for patch $x_{i,j}$ is given by $u_{i,j} = \frac{\mathcal{F}(x_{i,j})}{\|\mathcal{F}(x_{i,j})\|}$. Therefore, we obtain the prompt embeddings set $W = \{w_c\}_{c=1,...,C}$ for all categories.

Defined the classifier bank as a collection of H candidate classifiers: $\Psi = \{\psi_1, \psi_2, ..., \psi_H\}$. Here, each candidate classifier ψ_h is capable of mapping patch embedding $u_{i,j}$ and the prompt embeddings set W to a patch score $S_{x_{i,j}}^{\Psi_h} = \psi_h(u_{i,j}, W)$. We demonstrate an effective detailed implementation for the classifier bank in Sec. 2.2. The patch scores are subsequently utilized for patch filtering. Specifically, each candidate classifier ψ_h elects a subset of patches for WSI X_i, denoted as $Bag_{X_i}^{\psi_h}$, which includes the top q patches with the highest scores:

$$Bag_{X_i}^{\psi_h} = \{x_{i,j} \mid x_{i,j} \in \overset{(q)}{argmax}\, S_{x_{i,j}}^{\psi_h}\},\ \forall \psi_h \in \Psi. \tag{1}$$

Subsequently, a bank-nominated set $Bag_{X_i}^{\Psi}$ for the WSI X_i is obtained by taking a union of these sets: $Bag_{X_i}^{\Psi} = \bigcup_{\psi_h \in \Psi} Bag_{X_i}^{\psi_h}$. This filtering process removes patches with limited significance by a consensus among all candidate classifiers.

Further, the meta-learner $\mathcal{M}$, structured as a two-layer perceptron, predicts the classifier weights for the H candidate classifiers based on a patch embedding $u_{i,j}$. Specifically, given a nominated patch $x_{i,j} \in Bag_{X_i}^{\Psi}$ for WSI X_i, we obtain a set of H classifier weights $\Lambda_{i,j} = \{\lambda^1_{x_{i,j}}, \lambda^2_{x_{i,j}}, ..., \lambda^H_{x_{i,j}}\}$ by:

$$\Lambda_{i,j} = \mathcal{M}(u_{i,j}). \tag{2}$$

And the set of nominated patch prediction $p_{X_i} = \{p_{x_{i,j}}\}_{x_{i,j} \in Bag^{\Phi}_{X_i}}$ is calculated via:

$$p_{x_{i,j}} = \sum_{h=1}^{H} \lambda^h_{x_{i,j}} \cdot S^{\psi_h}_{x_{i,j}}. \tag{3}$$

Finally, we use a top-K max-pooling operation, h_{topK}, to get the WSI-level prediction $\mathcal{P}_{X_i}$ for WSI X_i:

$$\mathcal{P}_{X_i} = h_{\text{topK}}(p_{X_i}) = \frac{1}{K}\left[\sum_{i=1}^{K} \tilde{p}^1_i, \sum_{i=1}^{K} \tilde{p}^2_i, ..., \sum_{i=1}^{K} \tilde{p}^C_i\right], \tag{4}$$

where $\tilde{p}^c_i$ is the *i-th* largest score values from patch-level prediction set p_{X_i} for category c. The parameter of the meta-learner is optimized with the cross-entropy loss between the WSI-level prediction and ground truth.

2.2 Classifier Bank Construction with diverse Classifiers

Candidate classifiers in the classifier bank aim to provide complementary diagnostic emphases, allowing the meta-learner to comprehensively recognize the significance of each patch. In this section, we demonstrate an effective way to construct the classifier bank with four candidate classifiers, denoted as $\Psi = \{\psi_p, \psi_\sigma, \psi_\Delta, \psi_\beta\}$. Each candidate classifier is illustrated as follows:

- **Confidence Peak Classifier** ψ_p evaluates maximal similarity by computing the cosine similarity between the patch embedding and each prompt embedding, formulated as

$$S^{\psi_p}_{x_{i,j}} = u^T_{i,j} W. \tag{5}$$

- **Normalized Certainty Classifier** ψ_σ identifies easy distinguishability by applying a softmax function σ to the cosine similarity between patch embeddings and prompt embeddings. This is defined as:

$$S^{\psi_\sigma}_{x_{i,j}} = \sigma(u^T_{i,j} W). \tag{6}$$

- **Divergence Extremum Classifier** ψ_Δ also evaluates easy discrimination by computing the similarity difference of the highest two categories, formulated as:

$$S^{\psi_\Delta}_{x_{i,j}} = max_{(1)}(u^T_{i,j} W) - max_{(2)}(u^T_{i,j} W). \tag{7}$$

- **Background Suppression Classifier** ψ_β measures the minimal irrelevance by calculating the negative similarity between a patch and the background tissues. Specifically, we additionally introduce four normal tissue types, i.e., `‘stromal tissue’`, `‘inflammatory tissue’`, `‘vascular tissue’`, and `‘necrotic tissue’`. Following the same process routine as foreground categories, we obtain the background prompts $\{t_c^\beta\}_{c=1,\dots,C^\beta}$ and background prompt embedding $W^\beta = \{w_c^\beta\}_{c=1,\dots,C^\beta}$, where C^β denotes the number of background tissue types. Then, the patch score is given by:

$$S_{x_{i,j}}^{\psi_\beta} = -\sum_{c=1}^{C^\beta} u_{i,j}^T w_c^\beta. \tag{8}$$

3 Experiments

Dataset. We comprehensively compare our proposed MOC with many state-of-the-art (SoTA) methods using two real-world datasets: TCGA-NSCLC and TCGA-RCC. TCGA-NSCLC consists of 1052 WSI slides for lung cancer subtyping, and TCGA-RCC consists of 937 WSI slides for kidney cancer subtyping. Each dataset is randomly split five times into the training, validation, and test sets. For TCGA-NSCLC, each fold consists of 50 validation samples and 200 test samples, whereas for TCGA-RCC, each fold contains 20 validation samples and 70 test samples. To simulate the few-shot learning scenario, we then randomly select k samples per category from the training set to construct the k-shot experimental setup.

Implementation and evaluation We employ the CONCH [10] pretraining as the backbone for both the image and text encoders. We use the CLAM [12] toolkit for WSI preprocessing, and we set the patch size to 224 for feature extraction. We use the same prompt ensemble scheme following MI-Zero [11]. MOC w/o. ($\mathcal{M}$, Ψ) in Tab. 1 and Tab. 2 is implemented following MI-Zero [11]. For a

Table 1: Few-shot results on TCGA-NSCLC dataset. The best results are in bold, and the second-best results are underlined.

Method	1 shot		2 shot		4 shot		8 shot	
	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)
MIL-based Methods								
CLAM-SB [12]	52.39$_{\pm 6.24}$	48.38$_{\pm 1.43}$	58.03$_{\pm 5.76}$	48.23$_{\pm 1.36}$	69.22$_{\pm 4.74}$	51.95$_{\pm 5.15}$	73.68$_{\pm 6.72}$	59.96$_{\pm 9.44}$
CLAM-MB [12]	58.75$_{\pm 9.81}$	50.57$_{\pm 2.78}$	65.40$_{\pm 4.41}$	61.87$_{\pm 2.68}$	71.92$_{\pm 4.14}$	52.43$_{\pm 3.15}$	79.69$_{\pm 6.34}$	66.63$_{\pm 2.48}$
TransMIL [15]	62.24$_{\pm 4.59}$	55.24$_{\pm 5.15}$	63.95$_{\pm 4.95}$	55.87$_{\pm 5.17}$	74.51$_{\pm 5.66}$	61.42$_{\pm 7.95}$	83.69$_{\pm 7.05}$	75.82$_{\pm 7.44}$
ViLa-MIL [16]	71.79$_{\pm 4.64}$	56.48$_{\pm 6.79}$	72.93$_{\pm 3.35}$	60.34$_{\pm 3.60}$	77.79$_{\pm 4.88}$	57.86$_{\pm 7.66}$	84.20$_{\pm 7.09}$	73.51$_{\pm 9.22}$
Few-shot VLFM-based Methods								
CoOp [21]	62.04$_{\pm 8.24}$	59.15$_{\pm 7.84}$	70.21$_{\pm 2.78}$	63.97$_{\pm 2.47}$	72.23$_{\pm 4.18}$	58.65$_{\pm 3.42}$	80.11$_{\pm 5.91}$	69.48$_{\pm 5.93}$
TOP [13]	54.76$_{\pm 6.95}$	51.23$_{\pm 4.09}$	65.79$_{\pm 2.87}$	57.48$_{\pm 3.33}$	72.67$_{\pm 9.00}$	65.90$_{\pm 7.51}$	79.43$_{\pm 8.30}$	71.02$_{\pm 7.52}$
†**MOC w/o. ($\mathcal{M}$,Ψ)**	85.00$_{\pm 1.63}$	**75.31$_{\pm 1.93}$**	85.00$_{\pm 1.63}$	**75.31$_{\pm 1.93}$**	85.00$_{\pm 1.63}$	**75.31$_{\pm 1.93}$**	85.00$_{\pm 1.63}$	75.31$_{\pm 1.93}$
Our MOC	**88.29$_{\pm 2.65}$**	73.95$_{\pm 3.73}$	**89.11$_{\pm 1.13}$**	74.26$_{\pm 5.21}$	**90.65$_{\pm 1.02}$**	68.57$_{\pm 3.80}$	**90.51$_{\pm 1.74}$**	**77.10$_{\pm 3.80}$**

† Report the average zero-shot results on all test sets.

Table 2: Few-shot results on TCGA-RCC dataset. The best results are in bold, and the second-best results are underlined.

Method	1 shot		2 shot		4 shot		8 shot	
	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)
MIL-based Methods								
CLAM-SB [12]	$74.05_{\pm 10.56}$	$46.03_{\pm 6.07}$	$77.97_{\pm 12.92}$	$48.68_{\pm 7.12}$	$90.77_{\pm 1.24}$	$66.98_{\pm 4.62}$	$95.20_{\pm 1.09}$	$83.89_{\pm 2.78}$
CLAM-MB [12]	$75.81_{\pm 13.37}$	$56.49_{\pm 14.22}$	$77.97_{\pm 14.35}$	$44.09_{\pm 5.91}$	$91.34_{\pm 2.45}$	$76.68_{\pm 5.16}$	$95.53_{\pm 1.30}$	$82.35_{\pm 3.39}$
TransMIL [15]	$81.49_{\pm 10.36}$	$46.25_{\pm 7.82}$	$84.17_{\pm 9.52}$	$63.05_{\pm 14.31}$	$93.50_{\pm 1.88}$	$80.21_{\pm 3.73}$	$96.07_{\pm 1.59}$	$80.77_{\pm 8.06}$
ViLa-MIL [16]	$82.55_{\pm 12.31}$	$59.33_{\pm 12.75}$	$82.92_{\pm 10.43}$	$63.83_{\pm 11.95}$	$93.15_{\pm 2.64}$	$79.02_{\pm 5.97}$	$95.61_{\pm 2.36}$	$\underline{84.07_{\pm 5.37}}$
Few-shot VLFM-based Methods								
CoOp [21]	$87.40_{\pm 2.51}$	$54.33_{\pm 3.39}$	$82.59_{\pm 15.14}$	$50.65_{\pm 9.37}$	$92.92_{\pm 1.85}$	$58.28_{\pm 3.39}$	$94.98_{\pm 1.77}$	$58.81_{\pm 3.61}$
TOP [13]	$67.60_{\pm 3.45}$	$34.70_{\pm 4.46}$	$75.73_{\pm 3.96}$	$43.85_{\pm 10.04}$	$72.39_{\pm 3.64}$	$47.64_{\pm 8.44}$	$72.02_{\pm 4.40}$	$51.44_{\pm 10.23}$
†**MOC w/o. ($\mathcal{M}$, Ψ)**	$\underline{96.03_{\pm 0.68}}$	$\underline{80.39_{\pm 2.33}}$	$\underline{96.03_{\pm 0.68}}$	$\underline{80.39_{\pm 2.33}}$	$\underline{96.03_{\pm 0.68}}$	$\underline{80.39_{\pm 2.33}}$	$\underline{96.03_{\pm 0.68}}$	$80.39_{\pm 2.33}$
Our MOC	$\mathbf{96.25_{\pm 1.41}}$	$\mathbf{84.34_{\pm 5.24}}$	$\mathbf{97.45_{\pm 0.72}}$	$\mathbf{90.02_{\pm 2.55}}$	$\mathbf{97.42_{\pm 0.41}}$	$\mathbf{89.39_{\pm 2.75}}$	$\mathbf{97.78_{\pm 0.40}}$	$\mathbf{91.74_{\pm 1.44}}$

† Report the average zero-shot results on all test sets.

fair comparison, we apply the same splits, backbone, preprocessing, and prompts to all the methods. The learning rate is set to $1e^{-3}$, q in Eq. 1 is set to 1000, and K in Eq. 4 is set to 150. Hyperparameters for baseline methods are following the original implementation. We respectively report the average Area Under the Curve (AUC) and Accuracy (ACC) with corresponding standard deviation (±).

Remarkable improvements in few-shot WSI classification. We compare the performance of our proposed MOC with many SoTA MIL-based methods and few-shot VLFM-based methods on the TCGA-NSCLC dataset as Tab. 1 and the TCGA-RCC dataset as Tab. 2. The results suggest our method achieves the best on both datasets among different few-shot settings. Specifically, on the TCGA-NSCLC dataset, our MOC surpasses the second-best performing few-shot WSI classification method by 26.25%, 18.9%, 17.98%, 10.4% in AUC for 1, 2, 4, 8-shot settings and also outperforms zero-shot baseline by at least 3.29%.

Ablation study As shown in Tab. 3(left), our proposed MOC achieves the highest performance (89.64%) compared to the other three methods, demonstrating incremental performance gains through the successive integration ψ_β, ψ_Δ, and ψ_σ into baseline ψ_p. Tab. 3(right) reveals that the method incorporating four classifiers achieves the highest performance among all configurations, highlighting a consistent performance improvement as the number of classifiers increases.

The comparative analysis in Tab 4 provides critical insights into classifier integration strategies. Given multiple classifiers, a naive summation yields only marginal improvements (merely 0.4% increase over baseline), while our meta-learner-based integration achieves a remarkable 4.64% AUC enhancement, effectively leveraging complementary classifiers for a more holistic understanding.

Qualitative visualization We further depict the visualizations in Fig. 3. We compare our method with CoOp [21], TOP [13], and MI-Zero [11]. We find our MOC obviously shows better results than other baselines.

Table 3: Comparison of methods with (left) varying classifier combinations and (right) increasing number of classifiers (considering all possible classifier combinations).

Method	ψ_p	ψ_σ	ψ_Δ	ψ_β	avg AUC(%)
MOC w/o. ($\mathcal{M}$, $\varPsi$)	✔	✗	✗	✗	85.00
Method x	✔	✗	✗	✔	87.63
Method y	✔	✗	✔	✔	88.95
MOC	✔	✔	✔	✔	89.64

#Classifiers	avg AUC(%)
1	85.00
2	86.59
3	88.82
4	89.64

Table 4: Performance comparison of our baseline, multiple classifiers fused with summation, and our proposed MOC.

Method	average AUC(%)
MOC w/o. ($\mathcal{M}$, $\varPsi$)	85.00
Multiple Classifiers (w. Summation)	85.40
MOC (w. Meta-learner)	89.64

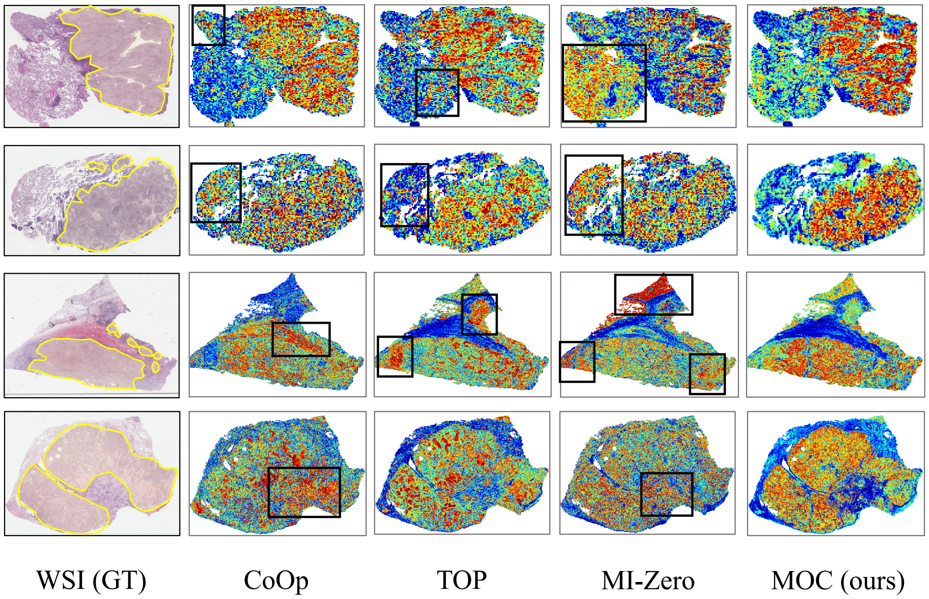


Fig. 3: Visualizations on the TCGA-NSCLC datasets. Ground-truth tumors are circled in yellow, predicted tumor regions are red, and black boxes note false positives.

4 Conclusion

In summary, this study presents a novel Meta-Optimized Classifier (MOC) for few-shot WSI classification. To address the limitation of the classifier design in existing VLFM-based few-shot explorations, we innovatively use a meta-learner to dynamically construct an optimal classifier from the classifier bank. Furthermore, we implement the classifier bank with diverse classifiers for a holistic pathological understanding. Extensive experiments demonstrate our MOC's state-of-the-art performance among multiple few-shot benchmarks.

Acknowledgments. This work is supported by a research grant from the National Natural Science Foundation of China (Grant No. 62306254) and a research grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. R6005-24).

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

References

1. Chikontwe, P., Kang, M., Luna, M., Nam, S., Park, S.H.: Low-shot prompt tuning for multiple instance learning based histology classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 285–295. Springer (2024)
2. Fu, K., Qu, L., Wang, S., Xiong, Y., Maglogiannis, I., Gao, L., Wang, M., et al.: Fast: A dual-tier few-shot learning paradigm for whole slide image classification. Advances in Neural Information Processing Systems **37**, 105090–105113 (2024)
3. Han, M., Qu, L., Yang, D., Zhang, X., Wang, X., Zhang, L.: Mscpt: Few-shot whole slide image classification with multi-scale and context-focused prompt tuning. IEEE Transactions on Medical Imaging (2025)
4. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
5. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
6. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021)
7. Li, H., Chen, Y., Chen, Y., Yu, R., Yang, W., Wang, L., Ding, B., Han, Y.: Generalizable whole slide image classification with fine-grained visual-semantic interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11398–11407 (2024)
8. Li, Y., Zhang, Q., Xiang, T., Lin, Y., Zhang, Q., Li, X.: Few-shot lymph node metastasis classification meets high performance on whole slide images via the informative non-parametric classifier. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–119. Springer (2024)
9. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19830–19839 (2023)
10. 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**(3), 863–874 (2024)

11. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19764–19775 (2023)
12. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
13. Qu, L., Fu, K., Wang, M., Song, Z., et al.: The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. Advances in Neural Information Processing Systems **36**, 67551–67564 (2023)
14. Qu, L., Yang, D., Huang, D., Guo, Q., Luo, R., Zhang, S., Wang, X.: Pathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification. In: European Conference on Computer Vision. pp. 196–212. Springer (2024)
15. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
16. Shi, J., Li, C., Gong, T., Zheng, Y., Fu, H.: Vila-mil: Dual-scale vision-language multiple instance learning for whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11248–11258 (2024)
17. Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple instance learning framework with masked hard instance mining for whole slide image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4078–4087 (2023)
18. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18802–18812 (2022)
19. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
20. Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V.B.: A graph-transformer for whole slide image classification. IEEE transactions on medical imaging **41**(11), 3003–3015 (2022)
21. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)