ViLa-MIL: Dual-scale Vision-language Multiple Instance Learning for Whole Slide Image Classification
—— Supplementary Materials ——

Jiangbo Shi[1], Chen Li[1*], Tieliang Gong[1], Yefeng Zheng[2], Huazhu Fu[3*]
[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
[2]Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China
[3]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
shijiangbo@stu.xjtu.edu.cn, cli@xjtu.edu.cn, hzfu@ieee.org

A. Dual-scale Visual Descriptive Text Prompt

The specific descriptions of dual-scale visual descriptive text prompts for renal cell carcinoma and lung cancer are shown in Figure S1 and Figure S2, respectively. Note that three experienced pathologists thoroughly examined the text prompts generated by GPT-3.5, and found them relatively correct and detailed.

B. Description of Datasets

The data statistics are reported in Table S1. The specific descriptions of our collected datasets are as follows:

- **TIHD-RCC**: This is the in-house dataset consisting of renal cell carcinoma slides collected by our research team. All the slides were stained with hematoxylin and eosin (H&E) and scanned by a KF-PRO-005 digital slice scanner at 20 $\times$ magnification with 0.5 μm/pixel resolution. There are 480 slides with slide-level subtyping labels.
- **TCGA-RCC**: To verify ViLa-MIL on multi-center data, we collected 639 renal cell carcinoma slides from TCGA.[1] The data in TIHD-RCC and TCGA-RCC are divided into three categories: clear cell (CCRCC), papillary (PRCC), and chromophore renal cell carcinoma (CRCC).
- **TCGA-Lung**: To verify the generalizability of ViLa-MIL across multiple cancer types, we also collected 658 lung cancer slides from TCGA. All the slides were annotated with slide-level subtyping labels. This dataset includes two subtypes: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC).

*Co-corresponding authors.
[1]https://portal.gdc.cancer.gov.

C. Implementation Details

Additional descriptions of implementation details are as follows. We adopt Adam optimization with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. The minimum training epoch number is 80, and the early stopping strategy is adopted if the accuracy does not continuously increase for 20 epochs. The batch size is 1. During training, all the methods utilize the same seed. ViLa-MIL and all the baselines are implemented with PyTorch and the PyG library [1] on a workstation with eight NVIDIA 2080Ti GPUs.

D. Description of Compared Methods

The specific description for each compared method is as follows:

- **Max-pooling**: Max-pooling is a baseline method that utilizes the max-pooling operator to generate the slide prediction.
- **Mean-pooling**: Mean-pooling is a baseline method that utilizes the mean-pooling operator to aggregate all the patch features as the slide features.
- **ABMIL [4]**: ABMIL proposes an attention-based aggregation method to generate the slide features by measuring the importance of each instance.
- **CLAM [7]**: CLAM proposes a global pooling operator trained for weakly-supervised slide-level classification tasks. CLAM-SB and CLAM-MB denote the single-attention-branch and multi-attention-branch versions of CLAM, respectively.
- **TransMIL [8]**: TransMIL proposes to utilize the self-attention mechanism to explore the global relations between patches.
- **DSMIL [5]**: DSMIL utilizes the multi-scale patches as the input and proposes a non-local attention-based

Cancer Type		Kidney		Lung
Dataset		TIHD-RCC	TCGA-RCC	TCGA-Lung
Number of WSIs		480	639	658
Data Split	Training	192	255	264
	Validation	144	192	197
	Test	144	192	197
Number of Patches	5×	429,402	607,817	490,977
	10×	1,670,362	2,359,471	1,903,894

Table S1. Data statistics.

fusion method.

- **GTMIL [10]**: GTMIL employs a graph representation to model the WSI data and utilizes a vision Transformer to generate slide features.
- **DTMIL [9]**: DTMIL proposes a double-tier MIL framework by introducing the concept of pseudo-bags.
- **IBMIL [6]**: IBMIL proposes an interventional training method based on the backdoor adjustment.

E. Comparisons with SOTA under Different Shots

To compare our ViLa-MIL with the current state-of-the-art WSI classification methods under different shots, we select four MIL-based methods (*i.e.*, CLAM-MB [7], TransMIL [8], GTMIL [10] and DTMIL [9]). The experiment results are summarized in Figure S3. As the number of shots increases, the performance of nearly all methods improves. Specifically, with fewer support samples (*i.e.*, 4-shot or 8-shot), our ViLa-MIL achieves more significant gains compared with all the other methods. This indicates that with the help of our dual-scale visual descriptive text prompt, ViLa-MIL can capture discriminative morphological patterns better for classification under the few-shot setting. With more support samples (*i.e.*, 32-shot or 64-shot), the performance of traditional MIL-based methods increases substantially; however, our ViLa-MIL still demonstrates better (or at least comparable) performance on all three datasets. Note that we have additionally conducted the experiment in a fully supervised setting. Specifically, compared to DTMIL, ViLa-MIL still achieves a comparable performance improvement of 0.8% and 0.7% in AUC and F1, respectively.

F. Interpretability Analysis

This section elaborates on the generation process of visualization results in Figure 5 of the main text. For the MIL-based models, we generate the visualization results by binarizing the attention maps. For our ViLa-MIL, to keep consistency with the other MIL-based methods, we only visualize the high-resolution visualization result. Specifically, the prototype-guided attention map (*i.e.*, $Q_h K_h^\top / \sqrt{d} \in \mathbb{R}^{N_h \times N_p}$) is first calculated. This map captures the cross-attention between patches and prototypes. Each row in the prototype-guided attention map represents a patch, while each column denotes a prototype. To establish the relationships between patches and prototypes, we assign each patch to the prototype with the highest cross-attention value. This grouping process ensures that each patch is associated with a specific prototype. Next, we employ the attention map A_h to select the prototype with the highest attention value. This prototype is deemed the most representative. Finally, we consider all the patches that are grouped into the cancer prototype as cancer patches, while the remaining patches are taken as normal. Note that to obtain the prediction result, MIL-based methods need to carefully select a threshold to binarize the attention map. For different WSIs, the best threshold may be different, it is hard to obtain the most optimal prediction result for each case. Our ViLa-MIL avoids this problem by utilizing the belonging relations between patches and prototypes based on the prototype-guided attention map.

G. The Setting of Each Module in Ablation Experiment

The module ablation settings in the ablation experiment are as follows:

- **ABMIL + Low-scale**: The attention-based method is utilized to aggregate the low-scale patch features as the baseline. Only the 5× patches and low-scale visual descriptive text prompt are utilized.
- **ABMIL + High-scale**: The attention-based method is utilized to aggregate the high-scale patch features as the baseline. Only the 10× patches and high-scale visual descriptive text prompt re utilized.
- **Patch Decoder + Low-scale**: Compared with the "ABMIL + Low-scale", the attention-based patch feature aggregation method is replaced with our proposed

Method	AUC	F1	ACC
Instance-level + Max Pooling	60.4±6.0	39.8±7.0	45.3±5.1
Instance-level + Top-K	74.9±6.4	57.0±7.6	59.4±7.6
Instance-level + Mean Pooling	78.0±2.5	60.4±5.4	61.2±5.3
ViLa-MIL	**84.3±4.6**	**68.7±7.3**	**68.8±7.3**

Table S2. Results (presented in %) of different similarity measurements on the TIHD-RCC dataset under 16-shot setting.

Method	AUC	F1	ACC
Feature Summation	79.3±2.9	62.7±4.3	65.1±2.3
Logit Summation	**84.3±4.6**	**68.7±7.3**	**68.8±7.3**

Table S3. Result (presented in %) of different multi-scale fusion methods on the TIHD-RCC dataset under 16-shot setting.

prototype-guided patch feature decoder.

- **Patch Decoder + High-scale**: Compared with the "ABMIL + High-scale", the attention-based patch feature aggregation method is replaced with our proposed prototype-guided patch feature decoder.
- **Patch Decoder + Multi-scale**: Compared with "Patch Decoder + Low-scale", the 10× patches and the high-scale visual descriptive text prompt are also utilized.
- **ViLa-MIL**: This is our complete framework proposed in this work. Compared with "Patch Decoder + Multi-scale", the context-guided text decoder is also introduced for both scales.

H. Effect of Text Prompt

To intuitively compare different text prompts, we present the visualization result of the "Class-name-replacement" template and our ViLa-MIL. As shown in Figure S4, compared with the "Class-name-replacement" template (the third column), our ViLa-MIL (the fourth column) achieves better localization results of the tumor. Since the "Class-name-replacement" template lacks the diagnosis-related prior, it easily fails to locate the tumor regions for subtyping with the supervision of few-shot labels. Our ViLa-MIL maintains a strong consistency with ground truth, which demonstrates that the dual-scale visual descriptive text prompt helps the model learn the discriminative morphological patterns from the WSI. Specifically, as shown in Figures S4(d) and S4(e), the low-scale visualization result locates the whole tumor structure well, and the high-scale visualization result presents more fine-grained details. From the patches with the highest similarity scores to CCRCC, we can observe that the low-scale patches show solid mass, well-circumscribed, and homogeneous texture, and the high-scale patches present clear cytoplasm, prominent nucleoli, and round or oval nuclei. Three experienced pathologists also confirm that these highlighted patches capture diagnosis-related patterns for each class.

I. Effect of Similarity Measurements

To verify the effect of different similarity measurements, we compare our bag-level similarity with several instance-level methods. Specifically, for the instance-level similarity calculation method, based on our ViLa-MIL, each patch feature directly calculates the similarity with the text prompt features, and then several aggregation methods are utilized to obtain the final slide-level similarity. As shown in Table S2, ViLa-MIL achieves the best results under all three metrics because the small patch contains limited visual information, which cannot match various kinds of text descriptions well.

J. Effect of Multi-scale Fusion Methods

To verify the effect of different multi-scale fusion methods, we compare our adopted logit summation with the feature summation method. In logit summation, the similarity between image and text features is calculated on each scale separately, and then the similarities of both scales are added together. Feature summation means that two-scale image and text features are summed together first, and then the similarity is calculated between the fused image and text features. As shown in Table S3, the logit summation method achieves better results compared with the feature summation method. Due to disparities in visual features at different scales and their corresponding textual descriptions, the feature summation method cannot achieve proper alignment for each scale, resulting in a decrease in performance.

K. Effect of Hyper-parameters

To verify the effect of hyper-parameters on the model's performance, we conduct a series of ablation studies on the TIHD-RCC dataset. The results are summarized in Figure S5. The best results are achieved with 16 prototypes (*i.e.*, $N_p = 16$). Too few prototypes cannot represent various kinds of visual features sufficiently. On the other hand, with too many prototypes, the redundant prototypes do not contribute to improving the model's performance but increase computational complexity. For the number of learnable text vectors M in the text prompt, the best value is 16. Fewer text vectors cannot effectively transfer the model to the pathological dataset. Conversely, employing more text vectors increases the number of parameters the model needs to learn. In the few-shot scenario, optimizing the model based on these limited data becomes challenging.

L. Effect of CLIP as the backbone

To verify the effect of domain-related VLMs on the model's performance, we replaced CLIP with PLIP and QuiltNet and conducted the experiments on the TIHD-RCC dataset in the 16-shot setting. Specifically, compared with CLIP (84.3%), PLIP (85.7%) [2] and QuiltNet (85.2%) [3] exhibit AUC improvements of 1.4% and 0.9 %. This indicates that VLMs pre-trained on the domain-specific data contribute to enhancing model performance further.

References

[1] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428*, 2019. 1

[2] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual-language foundation model for pathology image analysis using medical Twitter. *Nature Medicine*, pages 1–10, 2023. 4

[3] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1M: One million image-text pairs for histopathology. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[4] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the International Conference on Machine Learning*, pages 2127–2136, 2018. 1

[5] Bin Li, Yin Li, and Kevin W Eliceiri. 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*, pages 14318–14328, 2021. 1

[6] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19839, 2023. 2

[7] Ming Y Lu, Drew FK 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(6):555–570, 2021. 1, 2

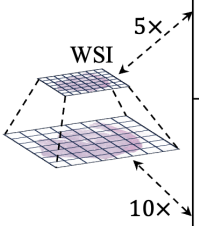
[8] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021. 1, 2

[9] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. 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*, pages 18802–18812, 2022. 2

[10] Yi Zheng, Rushin H. Gindra, Emily J. Green, Eric J. Burks, Margrit Betke, Jennifer E. Beane, and Vijaya B. Kolachalama. A graph-Transformer for whole slide image classification. *IEEE Transactions on Medical Imaging*, 41(11):3003–3015, 2022. 2

Question: What are the visually descriptive characteristics of {**class name**} at low and high resolution in the whole slide image?

LLM

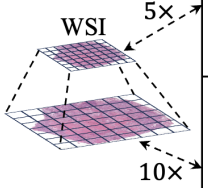


Class Name	Clear Cell Renal Cell Carcinoma	Papillary Renal Cell Carcinoma	Chromophobe Renal Cell Carcinoma
Low-scale Visual Descriptive Text Prompt	A whole slide image of clear cell renal cell carcinoma at low resolution with visually descriptive characteristics of solid mass, well-circumscribed borders, and homogeneous texture.	A whole slide image of papillary renal cell carcinoma at low resolution with descriptive characteristics of papillary growth pattern, multiple ures, well-circumscribed borders, varied architectural patterns, and pinkish coloration.	A whole slide image of chromophobe renal cell carcinoma at low resolution with visually descriptive characteristics of solid or well-circumscribed mass, homogeneous texture, distinct cell borders, and perinuclear clearing.
High-scale Visual Descriptive Text Prompt	A whole slide image of clear cell renal cell carcinoma at high resolution with visually descriptive characteristics of clear cytoplasm, round or oval nuclei, prominent nucleoli, rich vascularity, irregular blood vessels, intratumoral septa, and heterogeneity.	A whole slide image of papillary renal cell carcinoma at high resolution with visually descriptive characteristics of papillary structures with fibrovascular cores, nuclei arranged in layers, nuclear pseudo inclusions, psammoma bodies, heterogeneous cytoplasm, and inflammatory infiltrates.	A whole slide image of chromophobe renal cell carcinoma at high resolution with visually descriptive characteristics of halo around nuclei, perinuclear vacuolization, multiple nucleoli, binucleation or multinucleation, finely granular chromatin, and mitotic figures rare or absent.

Figure S1. Dual-scale visual descriptive text prompt for the renal cell carcinoma. By replacing the placeholder "class name" with the specific category label, such as Clear Cell Renal Cell Carcinoma, Papillary Renal Cell Carcinoma, and Chromophobe Renal Cell Carcinoma, the frozen LLM can generate the dual-scale visual descriptive text prompt that corresponds to the multi-scale WSIs.

Question: What are the visually descriptive characteristics of {**class name**} at low and high resolution in the whole slide image?

LLM



Class Name	Lung Adenocarcinoma	Lung Squamous Cell Carcinoma
Low-scale Visual Descriptive Text Prompt	A whole slide image of lung adenocarcinoma at low resolution with visually descriptive characteristics of irregularly shaped tumor regions, distortion or displacement of adjacent structures due to tumor growth, and smoothness or irregularity of tumor margins indicating invasiveness.	A whole slide image of lung squamous cell carcinoma at low resolution with visually descriptive characteristics of irregularly shaped tumor regions, often forming solid masses, distortion or displacement of adjacent structures due to tumor growth.
High-scale Visual Descriptive Text Prompt	A whole slide image of lung adenocarcinoma at high resolution with visually descriptive characteristics of clear cytoplasm, round or oval nuclei, prominent nucleoli, rich vascularity, irregular blood vessels, intratumoral septa, and heterogeneity.	A whole slide image of lung squamous cell carcinoma at high resolution with visually descriptive characteristics of squamous cell differentiation, round structures with eosinophilic cytoplasm, distinct cell borders, abundant cytoplasm, enlarged nuclei, irregular nuclear shape, and increased chromatin density.

Figure S2. Dual-scale visual descriptive text prompt for the lung cancer. By replacing the placeholder "class name" with the specific category label, such as Lung Adenocarcinoma and Lung Squamous Cell Carcinoma, the frozen LLM can generate the dual-scale visual descriptive text prompt that corresponds to the multi-scale WSIs.

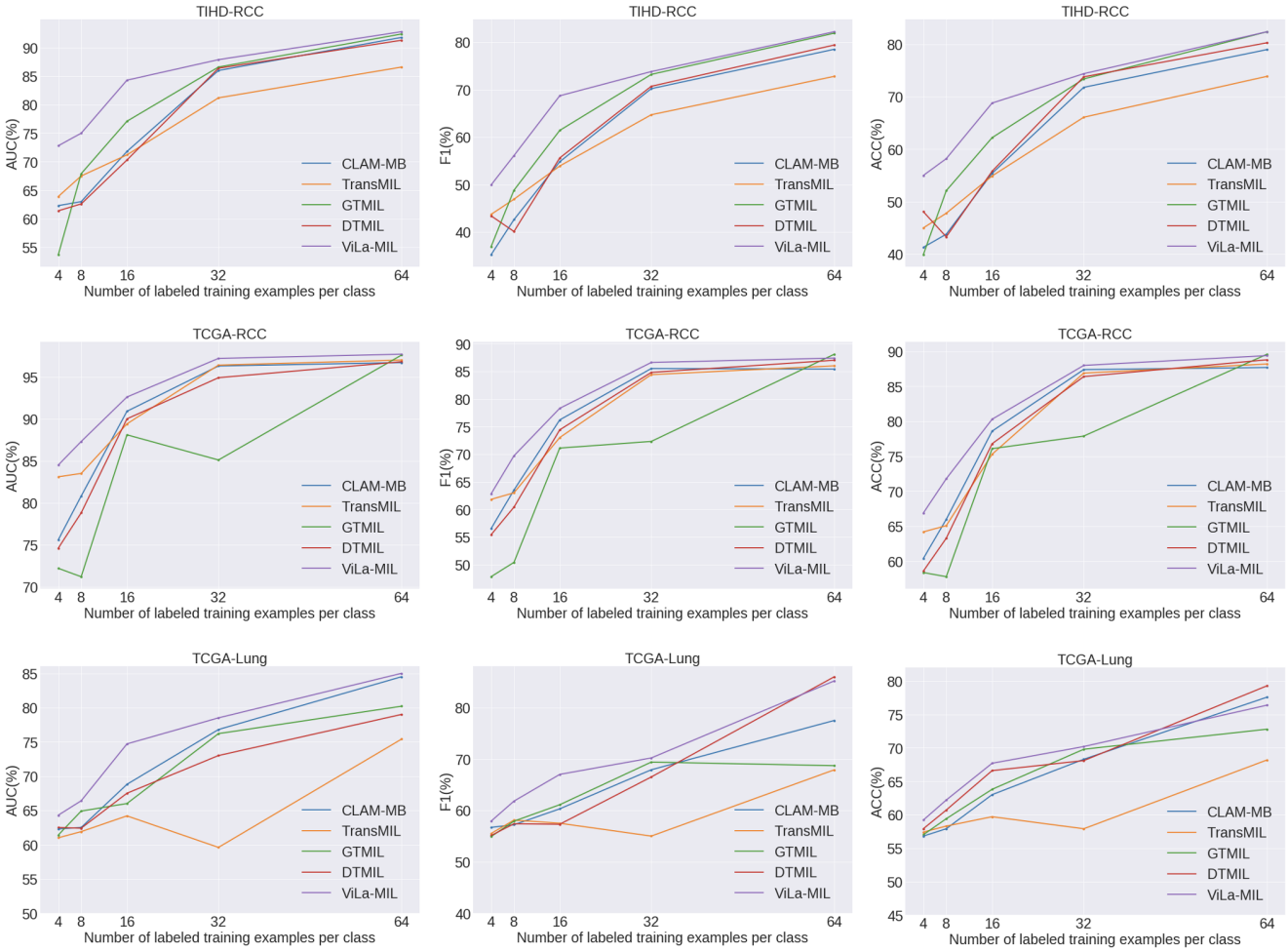


Figure S3. Performance comparison with different shots (4-/8-/16-/32-/64-shot) on TIHD-RCC, TCGA-RCC, and TCGA-lung datasets. N-shot denotes that each class has N training samples.

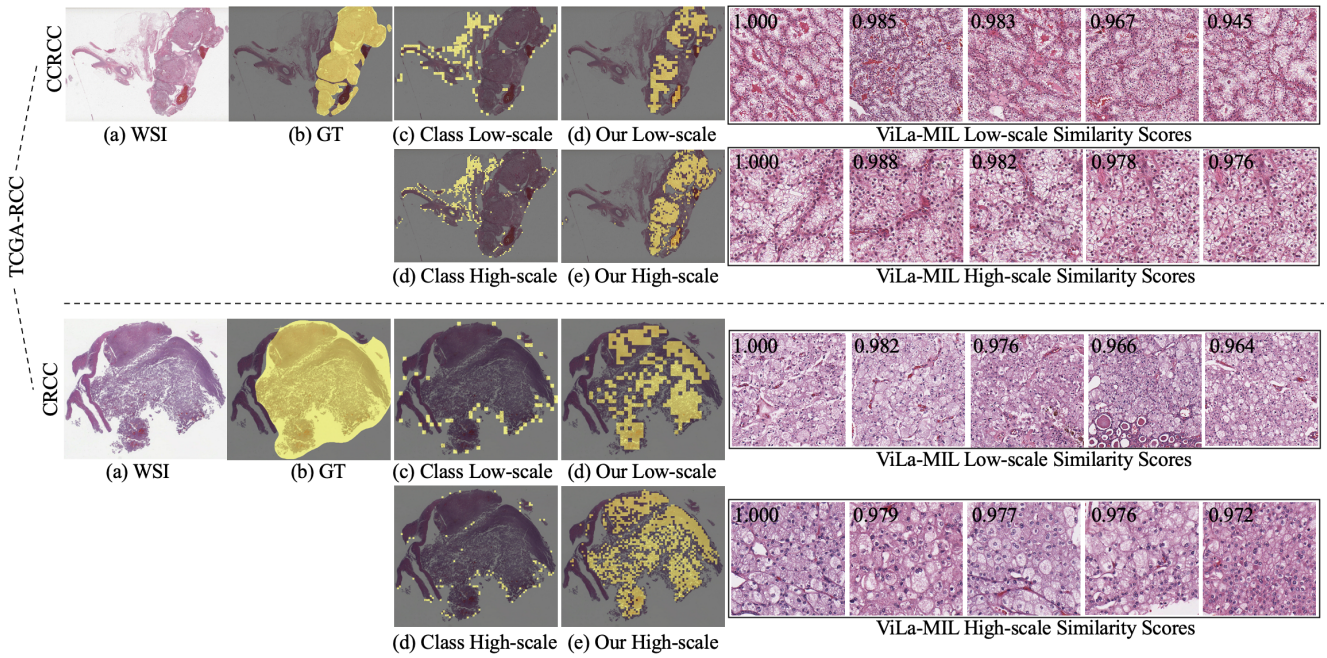


Figure S4. Comparison of our dual-scale visual descriptive text prompt with the class-name-replacement text prompt. Two cases (*i.e.*, CCRCC and CRCC) are randomly selected from the TCGA-RCC dataset to show the results of different text prompts. For each case, (a) is the original WSI; (b) is the corresponding ground truth (GT) tumor annotation; (c) and (d) are the visualization results by utilizing the "Class-name-replacement" template at low- and high-scale, respectively; (d) and (e) are the visualization results by utilizing our dual-scale visual descriptive text prompt at low- and high-scale, respectively. In the right half of the image, patches with the highest similarity are also visualized at low- and high-scale, respectively.

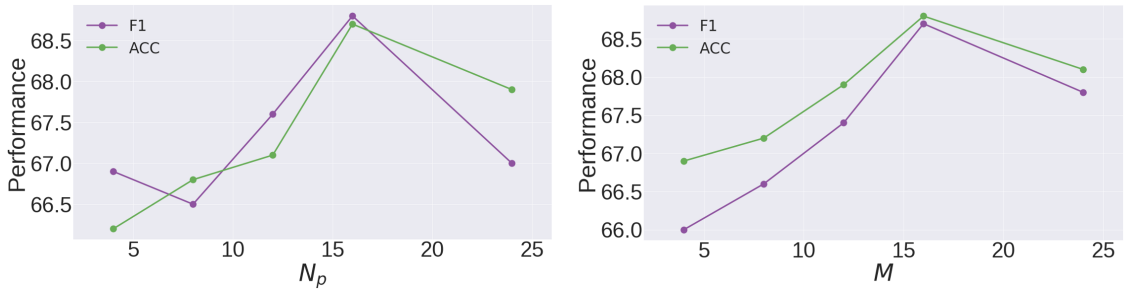


Figure S5. Impact of hyper-parameters: the number of prototypes N_p (left) and the number of learnable vectors M (right).