Non-Parametric Prototypes Enable Semantic Graph Learning for Whole-Slide Pathology

Zixuan Gao, Qing Zhang(✉), Hang Guo, Yan Wang, and Qingli Li

Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China
qzhang@cee.ecnu.edu.cn

Abstract. Placental pathological image classification plays a pivotal role in revealing pathological mechanisms underlying pregnancy-related diseases and supporting reliable clinical diagnosis. Existing whole slide image (WSI) classification methods often struggle to capture the substantial visual diversity across WSIs within a dataset, which limits their generalization ability. Additionally, current vision-language models lack effective semantic alignment between patch-level visual features and global contextual information, leading to insufficient fusion of visual and semantic cues. To address these challenges, we propose **ProtoSG-Net**, a **Proto**type-guided **S**emantic **G**raph learning network for WSI classification. Specifically, we design a non-parametric prototype generation module to model dataset-level representative visual diversity, guiding the key patch selection. Subsequently, a prototype-guided semantic calibration module is introduced to learn the relationships between selected key patches. The calibration module refines the semantic hierarchy of pathological attribute texts and further enhances the alignment between visual and textual features. Extensive experiments are conducted on a self-constructed placental dataset and 2 public datasets, demonstrating the effectiveness and efficiency of the proposed method. Code is available at https://github.com/ECNU-MultiDimLab/ProtoSG-Net.

Keywords: Whole Slide Image Classification · Multimodal Learning · Placental Pathology · Prototype Learning.

1 Introduction

Placental pathology assessment is essential for understanding the underlying mechanisms of pregnancy-related diseases and providing accurate clinical diagnosis [24, 5]. Accurate whole slide image (WSI) classification is fundamental in this process, revealing critical insights into tissue abnormalities by analyzing pathological information at various scales [6, 1]. Existing WSI classification methods select representative patches for efficient analysis. These patch selection methods tend to focus on individual WSIs without considering dataset-level

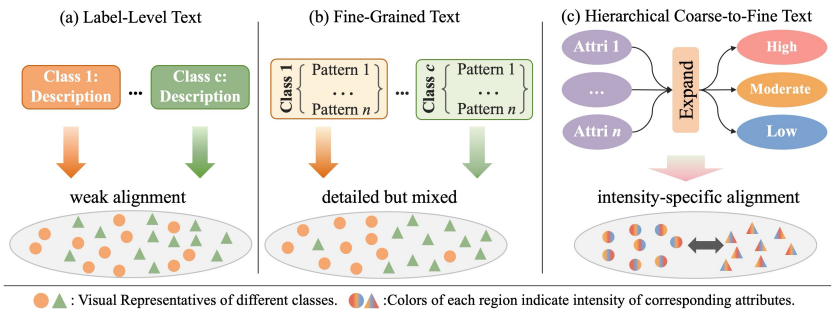

Fig. 1. Comparison of different textual prompts for vision-language alignment.

visual diversity [7, 21]. This diversity arises from variations like integrity, maturation, and inflammation, limiting the model's generalization ability and adversely affecting performance in real-world scenarios.

Textual prompts have been introduced to understand broader dataset context, which can be categorized into two main types: label-level text and fine-grained text [19, 10]. As shown in Fig. 1, the former one offers high-level disease-related pathological features but lacks the necessary details of WSIs [5, 15]. The latter one provides more detailed insights by describing various pathological patterns of each disease type [16, 6]. While this more detailed information is beneficial, it often treats each description as an isolated entity, failing to model how finer details build upon broader categories. Therefore, fine-grained text prompts struggle to capture hierarchical relationships between disease characteristics at different levels.

Since fine-grained text prompts are increasingly constructed for WSI analysis, the challenge lies in effectively integrating them with visual features. Existing vision-language fusion methods can be generally categorized into three approaches: contrastive learning [12, 4], cross-attention [5, 19, 27], and distance-based matching approaches [6, 20]. Contrastive learning-based methods maximize the similarity between matching visual-textual pairs but often struggle to capture the intricate relationships across modalities. Cross-attention mechanisms are designed to align visual and textual features, but they are prone to overfitting and fail to handle complex hierarchical relationships. Distance-based matching, while useful for measuring similarity, often lacks the flexibility needed to model the full diversity of visual and textual data. While these methods have been successful in several tasks, they face limitations in addressing dataset-level diversity and complex hierarchical structures.

To solve these challenges, we propose a **Proto**type-guided **S**emantic **G**raph learning **Net**work for WSI classification, named as a **ProtoSG-Net**. Specifically, we introduce an attribute-aware prototype guided patch selection module, leveraging the non-parametric dataset-level prototype generated from WSIs and the coarse-grained attribute texts. Additionally, we design a semantic-aware prototype guided graph learning module, which models the interactions between

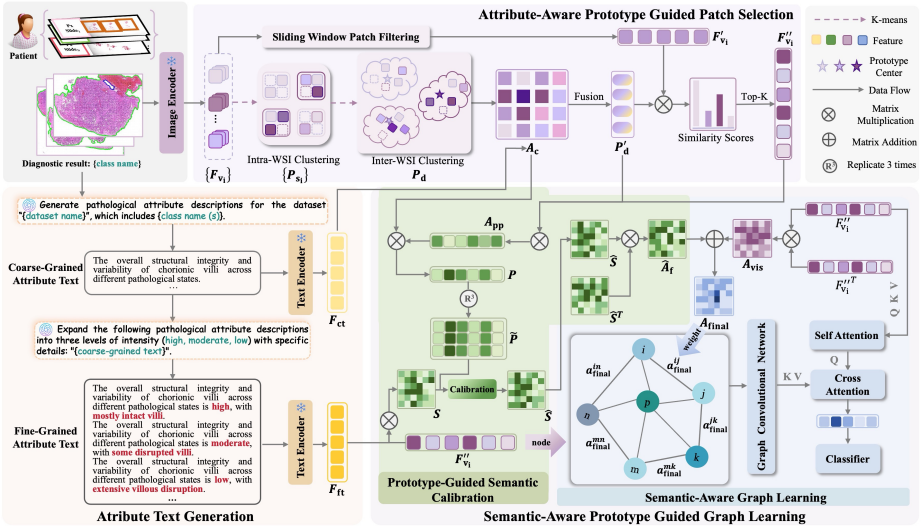

Fig. 2. Overview of the proposed ProtoSG-Net: Key Patches are selected in *Attribute-Aware Prototype-Guided Patch Selection* module guided by the dataset-level prototypes generated based on coarse-grained texts. The relationship between adjacent patches is learned in *Semantic-Aware Prototype-Guided Graph Learning* module. Specifically, the semantic adjacency information derived from fined-grained multimodal feature is calibrated by the former prototypes. Finally, the semantic information guides the global feature analysis by graph learning for precise WSI classification.

fine-grained attributes and key patches, and calibrates them using the above prototypes. The calibrated semantic feature enables more effective analysis of patch correlations beyond spatial adjacency, improving WSI classification.

2 Method

2.1 Problem Definition and Network Structure

Given a pathological WSI dataset $\mathcal{D} = \{\mathbf{X}, \mathbf{T}, \mathbf{Y}\}$ where $\mathbf{X}$ is the WSI set, $\mathbf{T} = \{\mathbf{T}_{\mathrm{c}}, \mathbf{T}_{\mathrm{f}}\}$ is its corresponding coarse- and fine-grained descriptions and $\mathbf{Y}$ is the true classification labels, our goal is to predict the disease type. Fig. 2 illustrates the architecture of the proposed ProtoSG-Net for WSI classification. For the i^{th} case, we extract its patch-level features using a pre-trained image encoder and construct non-parametric visual prototypes to capture dataset-level morphological diversity and guide the selection of the top-K key patches. To enhance multimodal alignment, we further incorporate pathological attribute texts into a semantic calibration module to obtain the relationship between adjacent patches. The relationship matrix guide the following graph learning for robust and generalizable case-level prediction.

2.2 Attribute-Aware Prototype Guided Patch Selection

Primary Multimodal Feature Extraction. For each patient, we partition the WSI into non-overlapping patches and extract patch-level features using the pre-trained pathology vision encoder UNI [2]. For the i^{th} case, patch embedding set is denoted as $F_{\mathrm{v}_i} = \{f_{\mathrm{v}_{ij}}\}_{j=1}^{N_i}$, where N_i represents the number of patches and $f_{\mathrm{v}_{ij}}$ is the j^{th} embedding. To reduce patch redundancy, we first apply an $n \times n$ sliding window over the patch grid to remove visually similar patches by computing pairwise similarities between patch embeddings [5]. In the placenta dataset, where each patient may correspond to multiple WSIs, we concatenate the features from all slides of the same patient as a unified patient-level WSI representation. This procedure produces a filtered patch-level feature set F'_{v_i}.
Non-Parametric Prototype Generation. To further capture representative visual features, we design a non-parametric prototype generation module by integrating intra- and inter-WSI information. For the i^{th} WSI, we perform intra-WSI K-means clustering on its patch-level features to derive slide-level prototypes P_{s_i}:

$$P_{\mathrm{s}_i} = \texttt{KMeans}(F_{\mathrm{v}_i}) = \{p_{\mathrm{s}_{ik}}\}_{k=1}^{K_{\mathrm{S}}}, \tag{1}$$

where $p_{\mathrm{s}_{ik}}$ denotes the k^{th} prototype of the i^{th} WSI and K_{S} is the number of the slide-level prototypes. While the prototypes can capture local morphological patterns, they are confined to individual WSI distributions. To model dataset-level visual diversity, we aggregate prototypes from all WSIs and perform inter-WSI clustering to derive global dataset-level prototypes P_{d}. These dataset-level prototypes summarize recurrent tissue patterns across slides and provide a compact representation of dataset-wide morphological variations.

$$P_{\mathrm{d}} = \texttt{KMeans}\left(\bigcup_{i=1}^{M} P_{\mathrm{s}_i}\right) = \{p_{\mathrm{d}_j}\}_{j=1}^{K_{\mathrm{D}}}, \tag{2}$$

where M is the number of cases, K_{D} is the number of dataset-level prototypes and p_{d_j} is the j^{th} prototype.

Beyond visual features, we also introduce coarse-grained attribute texts to provide high-level semantic descriptions of tissue morphology, encompassing key pathological properties such as tissue differentiation, spatial organization, cellular morphology, and boundary characteristics. These texts serve as global semantic priors of subsequent visual -semantic modeling. These attribute texts are encoded using a pre-trained text encoder BERT [3] to obtain textual feature $F_{\mathrm{ct}} = \{t_i \mid i = 1, \cdots, N_{\mathrm{T}}\}$, where N_{T} is the number of attribute texts. Subsequently, we project the textual feature F_{ct} into the visual embedding space via a fixed transformation, yielding $\tilde{t}_i$, and compute the cosine similarity between each visual prototype $p_{\mathrm{d}_i} \in P_{\mathrm{d}}$ and the projected textual feature F_{ct} to construct a visual-semantic association matrix $\mathbf{A}_{\mathrm{c}} \in \mathbb{R}^{K_{\mathrm{D}} \times N_{\mathrm{T}}}$. This association matrix is then utilized to aggregate semantic information for prototype refinement.

$$A_{\mathrm{c}_{ij}} = \frac{p_{\mathrm{d}_i}^{\top} \tilde{t}_j}{\|p_{\mathrm{d}_i}\|_2 \, \|\tilde{t}_j\|_2}, \quad \hat{t}_i = \sum_{j=1}^{N_{\mathrm{T}}} \mathrm{Softmax}(A_{\mathrm{c}_{ij}})\, \tilde{t}_j. \tag{3}$$

The aggregated semantic representation is concatenated with the original visual prototypes to obtain the refined prototype set:

$$P'_{\mathrm{d}} = \{p'_{\mathrm{d}_i}\}_{i=1}^{K_{\mathrm{D}}} = \left\{p_{\mathrm{d}_i} \parallel \hat{t}_i\right\}_{i=1}^{K_{\mathrm{D}}}. \tag{4}$$

The refined dataset-level prototypes are then projected into the same feature space as the patch embeddings. We compute the pairwise similarities between them and perform a second-stage selection by retaining the top-N patches with the highest similarity scores. The filtered feature set is defined as $F''_{\mathrm{v}_i} = \{f_{\mathrm{v}_{ij}}\}_{j=1}^{N}$.

2.3 Semantic-Aware Prototype-Guided Graph Learning

Prototype-Guided Semantic Calibration. To bridge the semantic granularity gap between global attribute descriptions and local patch-level representations, each coarse-grained pathological attribute is expanded into three fine-grained semantic states with different intensity levels. These fine-grained attribute texts are encoded using a pre-trained BERT model, producing textual embeddings $F_{\mathrm{ft}} = \{\tilde{t}_i \mid i = 1, \cdots, 3N_{\mathrm{T}}\}$. Subsequently, we compute patch-text similarities between F_{ft} and the filtered patch features F''_{v_i} in a unified embedding space, generating a patch-text affinity matrix $\mathbf{S}$. This affinity is then propagated through the shared attribute space to derive a semantic adjacency matrix: $\mathbf{A}_{\mathrm{f}} = \mathbf{S}\mathbf{S}^{\top}$, where each entry quantifies the semantic correlation between two patches mediated by fine-grained pathological attributes.

Patch–text similarities derived from fine-grained attributes capture only local semantics and may suffer from noise due to limited contextual information. To address this issue, we introduce a prototype-guided semantic calibration module, where global prototypes serve as dataset-level intermediates encoding the semantic association of recurrent visual patterns. Specifically, we compute the similarity between the refined prototypes P'_{d} and the final filtered patch features F''_{v_i}, generating a patch-prototype similarity matrix $\mathbf{A}_{\mathrm{pp}} \in \mathbb{R}^{N \times K_{\mathrm{D}}}$. We reuse the prototype–attribute similarity matrix $\mathbf{A}_{\mathrm{c}}$ to obtain a coarse-grained patch-attribute preference matrix: $\mathbf{P} = \mathbf{A}_{\mathrm{pp}}\mathbf{A}_{\mathrm{c}}$, where $\mathbf{P} \in \mathbb{R}^{N \times N_{\mathrm{T}}}$. To align with fine-grained attribute modeling, each column of $\mathbf{P}$ is replicated three times to form a fine-grained patch-attribute preference matrix $\tilde{\mathbf{P}} \in \mathbb{R}^{N \times 3N_{\mathrm{T}}}$. While $\mathbf{P}$ captures coarse attribute preferences at a global level, $\tilde{\mathbf{P}}$ encodes fine-grained semantic priorities for each patch under the prototype-guided prior.

We then calibrate the patch-text similarity matrix $\mathbf{S}$ by incorporating the prototype-guided preference: $\hat{\mathbf{S}} = \gamma\tilde{\mathbf{P}} + \mathbf{S}$, where γ is a learnable scaling coefficient controlling the strength of semantic calibration. The calibrated semantic adjacency matrix is defined as: $\hat{\mathbf{A}}_{\mathrm{f}} = \hat{\mathbf{S}}\hat{\mathbf{S}}^{\top}$. Finally, to jointly model semantic and visual relationships among patches, we construct the unified adjacency matrix:

$$\mathbf{A}_{\mathrm{final}} = \alpha\hat{\mathbf{A}}_{\mathrm{f}} + (1-\alpha)\mathbf{A}_{\mathrm{vis}}, \tag{5}$$

where $\mathbf{A}_{\mathrm{final}} \in \mathbb{R}^{N \times N}$, $\alpha \in [0, 1]$ balances semantic and visual contributions, and $\mathbf{A}_{\mathrm{vis}}$ denotes visual adjacency matrix computed from patch-level visual features.

Semantic-Aware Graph Learning. Given the final adjacency matrix $\mathbf{A}_{\text{final}}$, we construct a patch-level graph that encodes inter-patch relationships, followed by a lightweight graph convolutional network (GCN). The GCN-generated features serve as the Key and Value in the subsequent cross-attention module in decoder. The filtered patch features F''_{v_i} are enhanced via a self-attention block to capture local contextual dependencies and used as the Query in the cross-attention. This cross-attention mechanism enables each patch to selectively attend to graph-enhanced representations, effectively integrating both local contextual information and global visual-semantic structure.

$$F_{\text{patch}_i} = \texttt{CrossAttn}(\texttt{SelfAttn}(F''_{\text{v}_i}), \mathbf{H}, \mathbf{H}), \mathbf{H} = \texttt{GCN}(\mathbf{A}_{\text{final}}, F''_{\text{v}_i}). \quad (6)$$

Finally, global mean pooling is applied across the patch dimension to derive a case-level representation for classification. For optimization, we adopt Focal Binary Cross-Entropy Loss [13]. For testing, the dataset-level prototypes generated from the training set are directly reused for patch selection, while all other inference procedures remain identical to those in training.

3 Experiments

3.1 Experimental Settings

Datasets. We evaluate ProtoSG-Net on three datasets: a private Placenta dataset and two public datasets (TCGA-NSCLC [23] and Camelyon+ [14]). Placenta contains 1,483 WSIs from 620 patients with three pathological subtypes: avascular villi (284), placental infarction (177), and acute chorioamnionitis (375). Each patient provides 2–4 WSIs, treated as a single bag for patient-level multi-label classification. The dataset is split at patient level into training, validation, and test sets with a 7:1.5:1.5 ratio. These public datasets are separately split using a 6:2:2 ratio. To construct hierarchical attribute texts for visual–semantic alignment, we use GPT-5.2 to generate pathology-related descriptions. Coarse-grained morphological attributes are defined and expanded into three intensity levels (high, moderate, low) to form fine-grained texts. All texts are manually reviewed for medical plausibility and semantic consistency.

Implementation Details. We use the Adam optimizer and an initial learning rate of 1×10^{-4}, which decays linearly during training. Patch filtering uses a sliding window of size 4×4. For Placenta and TCGA-NSCLC, K_{S} is set to 7, K_{D} to 15, and N to 4000. For Camelyon+, K_{S} is set to 20, K_{D} to 15, and N to 3000. We set $\alpha = 0.5$ and initialize the learnable parameter γ to 0.3. To evaluate performance at both patient and slide level, we report Acc, F1, AUROC and PRAUC, which are suitable for multi-label and class-imbalanced classification tasks. All experiments are conducted on one NVIDIA GeForce RTX 4090 GPU.

3.2 Comparison with SOTA Methods

As shown in Table 1, we compare ProtoSG-Net with current state-of-the-art models. EmmPD [5] is a vision-language model, while the others are purely

Table 1. Classification performance (%) evaluation of our model and SOTA methods. The best results are highlighted in bold, and the second-best results are underlined.

Method	Placenta				TCGA-NSCLC				Camelyon+			
	ACC	F1	ROC AUC	PR AUC	ACC	F1	ROC AUC	PR AUC	ACC	F1	ROC AUC	PR AUC
ABMIL[8]	65.23	67.56	70.95	61.72	94.74	94.63	98.68	98.83	86.52	62.82	91.66	70.96
CLAMSB[17]	74.19	71.88	77.97	71.36	95.69	95.48	99.17	99.23	87.59	65.13	88.44	68.71
TransMIL[18]	67.03	61.98	69.77	61.72	96.17	96.00	99.14	99.28	88.65	65.12	92.85	74.47
DSMIL[9]	45.88	62.90	67.59	63.77	95.22	95.10	99.08	99.18	86.52	66.05	88.98	69.92
DTFDMIL[26]	45.88	62.90	75.00	68.95	94.74	94.53	98.26	98.41	87.23	62.21	89.31	69.21
S4MIL[25]	70.25	58.29	67.60	61.72	95.69	95.65	99.12	99.17	87.94	66.65	91.26	73.46
WiKG[11]	73.84	70.45	75.73	69.24	95.22	95.15	98.41	98.28	86.88	60.23	85.57	68.08
RRTMIL[22]	64.16	57.98	60.60	57.30	95.69	95.69	98.94	99.09	88.30	64.78	91.03	74.90
GDFMIL[28]	69.53	64.14	71.79	66.87	94.74	94.79	99.29	99.39	87.23	62.56	87.95	69.23
GMamba[29]	66.31	60.17	69.13	62.69	93.78	93.66	98.31	98.36	87.23	62.79	86.26	66.62
EmmPD[5]	76.34	72.95	**79.36**	72.15	94.74	94.74	99.13	99.19	86.88	66.02	92.02	70.55
ProtoSG-Net	**78.49**	**75.61**	78.70	**73.46**	**97.13**	**97.00**	**99.36**	**99.45**	**89.36**	**72.66**	**92.86**	**77.35**

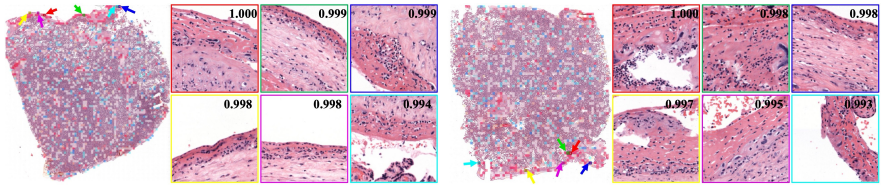

Fig. 3. Attention heatmaps of two slides from one case, showing the retained patches with the highest attention scores highlighted.

.

Table 2. Ablation study (%) of key components in ProtoSG-Net.

Module	Placenta				TCGA-NSCLC				Camelyon+			
	ACC	F1	ROC AUC	PR AUC	ACC	F1	ROC AUC	PR AUC	ACC	F1	ROC AUC	PR AUC
w/o SC&GCN	70.97	67.21	75.66	66.56	94.74	94.74	99.32	99.41	84.75	60.81	86.94	68.13
w/o SC&Text	72.40	68.83	74.94	67.97	95.22	95.05	99.13	99.18	86.17	60.64	90.11	68.65
w/o SC	75.27	69.33	76.41	69.36	96.17	96.12	99.19	99.35	87.94	66.03	90.79	69.31
w/o PS	73.84	66.36	75.44	69.65	96.17	96.12	98.85	99.05	86.52	61.23	89.21	68.90
ProtoSG-Net	**78.49**	**75.61**	**78.70**	**73.46**	**97.13**	**97.00**	**99.36**	**99.45**	**89.36**	**72.66**	**92.86**	**77.35**

image-based models. All models use UNI [2] as the feature extractor. Under identical settings, ProtoSG-Net consistently outperforms all baselines across three datasets. In particular, ProtoSG-Net improves Accuracy, F1, AUROC, and PRAUC by 0.71-32.61%, 1.00-17.63%, 0.01-18.1%, 0.06-16.16%, respectively. To further validate these results, we visualize two WSIs from a representative placenta case in Fig. 3. We generate attention heatmaps over the patches retained

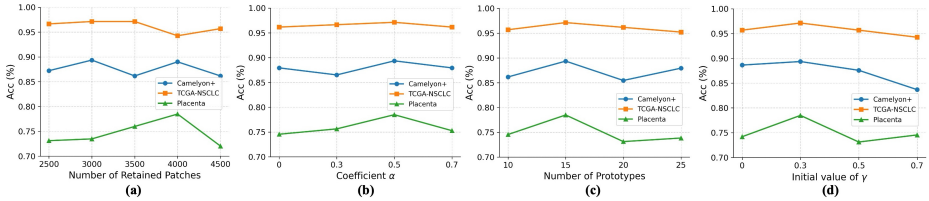

Fig. 4. Impact of Hyper-Parameters.

by the patch selection module and highlight the top-scoring patches, with their attention scores annotated in the upper-right corner. These high-attention regions are concentrated in the typical lesion areas of acute chorioamnionitis. The corresponding patches show neutrophil infiltration and enhanced eosinophilic staining, demonstrating that the model focuses on pathologically meaningful regions. All the results demonstrate the superiority of the proposed ProtoSG-Net.

3.3 Ablation Study

Impact of Key Modules. We conduct ablation studies by progressively removing key modules from the full model, including prototype-guided patch selection (PS), semantic calibration (SC), the GCN module, and the text branch. Specifically, removing SC disables semantic calibration on $\mathbf{A}_{\mathrm{f}}$, while the graph is still constructed based on $\mathbf{A}_{\mathrm{final}}$. Removing the text branch while preserving the GCN forces graph learning to rely solely on $\mathbf{A}_{\mathrm{vis}}$, with prototypes generated without textual information. As shown in Table 2, removing the prototype-related components consistently leads to performance degradation, highlighting the importance of non-parametric prototypes in capturing representative patterns and the importance of semantic calibration in constructing reliable semantic graphs.

Impact of Hyperparameters. To investigate the impact of key hyperparameters on model performance, we conduct experiments by varying number of retained patches (N), coefficient α, number of prototypes (K_{D}), and initial value of γ, as shown in Fig. 4. The optimal numbers of patches for Placenta, TCGA-NSCLC, and Camelyon+ are 4000, 3500, and 3000, respectively. A coefficient α of 0.5 provides the best results across all datasets, balancing semantic and visual alignment for graph construction. The number of prototypes is set to 15 for optimal performance, as increasing the number leads to overfitting. The learnable parameter γ achieves the best performance when initialized to 0.3.

Text Robustness and Fairness Analysis. We further evaluate the reliability of the textual branch from two aspects. For robustness, we add Gaussian noise (mean = 0, std = 0.01, 0.1, 0.5) to text embeddings, and observe that performance remains stable with less than 2% variation in Acc and F1 across all datasets. For a fair comparison, EmmPD [5], a multimodal model with textual inputs in Section 3.2, is evaluated under the same GPT-5.2-generated text setting as our method. Under this setting, EmmPD achieves 75.27%, 94.74%, 87.59% Acc on Placenta, TCGA-NSCLC, and Camelyon+, respectively, while

our method consistently outperforms it across all datasets, indicating that the improvements are not due to differences in LLM backbones.

4 Conclusion

In this work, we propose a prototype-guided semantic graph learning network for WSI classification. By introducing an attribute-aware prototype-guided patch selection module and a semantic-aware graph learning mechanism, our method strengthens multimodal alignment between hierarchical attributes and key visual patches, thereby enabling effective graph learning. Extensive experiments on three datasets exhibit exceptional performance and strong generalization capability of our method. In future work, we plan to validate our method on larger multi-center datasets and explore its extension to other downstream works.

Acknowledgments. This work is supported by National Natural Science Foundation of China (Grant No. 62475072, 42530110), Fundamental Research Funds for the Central Universities, the Science and Technology Commission of Shanghai Municipality (Grant No. 25xtcx00600, 22DZ2229004), Joint Laboratory of Hyperspectral Big Data and Artificial Intelligence (No. MIP202602), the AI project from the economic and information commission of Shanghai (Grant No. 2024-GZL-RGZN-01038).

Disclosure of Interests. The authors have no competing interests to declare in relation to this paper.

References

1. Bontempo, G., Bolelli, F., Porrello, A., Calderara, S., Ficarra, E.: A graph-based multi-scale approach with knowledge distillation for wsi classification. IEEE Transactions on Medical Imaging **43**(4), 1412–1421 (2023)
2. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
4. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., et al.: A multimodal whole-slide foundation model for pathology. Nature medicine pp. 1–13 (2025)
5. Guo, H., Zhang, Q., Gao, Z., Yang, S., Peng, S., Tao, X., Yu, T., Wang, Y., Li, Q.: Efficient multi-slide visual-language feature fusion for placental disease classification. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 8018–8027 (2025)
6. 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)

7. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2424–2433 (2016)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
9. 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)
10. 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)
11. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11323–11332 (2024)
12. Liang, Y., Lyu, X., Chen, W., Ding, M., Zhang, J., He, X., Wu, S., Xing, X., Yang, S., Wang, X., et al.: Wsi-llava: A multimodal large language model for whole slide image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22718–22727 (2025)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
14. Ling, X., Lei, Y., Li, J., Cheng, J., Huang, W., Guan, T., Guan, J., He, Y.: Towards a comprehensive benchmark for pathological lymph node metastasis in breast cancer sections. arXiv preprint arXiv:2411.10752 (2024)
15. Ling, X., Ping, Y., Li, J., Peng, J., Chen, Y., Ouyang, M., Wang, Y., He, Y., Guan, T., Liu, X., et al.: Multimodal distillation-driven ensemble learning for long-tailed histopathology whole slide images analysis. arXiv preprint arXiv:2503.00915 (2025)
16. Liu, P., Ji, L., Gou, J., Fu, B., Ye, M.: Interpretable vision-language survival analysis with ordinal inductive bias for computational pathology. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=trj2Jq8riA
17. 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)
18. 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)
19. 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)
20. Sun, Y., Si, Y., Zhu, C., Gong, X., Zhang, K., Chen, P., Zhang, Y., Shui, Z., Lin, T., Yang, L.: Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10360–10371 (2025)

21. Sun, Y., Wu, H., Zhu, C., Si, Y., Chen, Q., Zhang, Y., Zhang, K., Li, J., Cai, J., Wang, Y., et al.: Pathbench: Advancing the benchmark of large multimodal models for pathology image understanding at patch and whole slide level. IEEE Transactions on Medical Imaging (2025)
22. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11343–11352 (2024)
23. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: Review the cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia **2015**(1), 68–77 (2015)
24. Vanea, C., Džigurski, J., Rukins, V., Dodi, O., Siigur, S., Salumäe, L., Meir, K., Parks, W.T., Hochner-Celnikier, D., Fraser, A., et al.: Mapping cell-to-tissue graphs across human placenta histology whole slide images using deep learning with happy. Nature Communications **15**(1), 2710 (2024)
25. Yang, S., Wang, Y., Chen, H.: Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. arXiv preprint arXiv:2403.06800 (2024)
26. 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)
27. Zhang, Q., Li, L., Cai, R., Li, Q., Dissanayake, B., Wang, Y., Kot, A.: Ps-net: high-frequency attention and bayesian analysis based facial pore segmentation with no human annotation. Visual Intelligence **3**(1), 20 (2025)
28. Zhang, Y.X., Zhou, Z., Liu, W., Zhang, M.: Rethinking multi-instance learning through graph-driven fusion: A dual-path approach to adaptive representation. In: AAAI (2026)
29. Zheng, T., Yao, H., Jiang, K., Xiao, Y., Zhao, S.: Gmmamba: Group masking mamba for whole slide image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9935–9944 (2025)