PathoPrompt: Cross-Granular Semantic Alignment for Medical Pathology Vision-Language Models

Runlin Huang[1], Haohui Liang[1], Hongmin Cai[3], Weipeng Zhuo[1,2], Wentao Fan[1,2], and Weifeng Su[1,2]*

[1] Division of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai 519087, China
[2] Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, Beijing Normal-Hong Kong Baptist University, Zhuhai 519087, China
[3] School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

Abstract. Pre-trained visual-language (V-L) models have demonstrated impressive generalization capabilities on various downstream tasks, yet their performance is significantly influenced by the input text prompts. Previous studies (e.g., CoPrompt) have attempted to use detailed descriptions generated by LLM to assist model learning. For example, while a coarse-grained prompt like "A photo of Debris." may be less informative, a fine-grained description such as "Debris consists of dead cells and matrix fragments." provides additional context, resulting in enhanced model performance. However, existing methods generally lack the sensitivity to capture the subtle semantic differences that are crucial for accurately classifying pathology images. To tackle this challenge, we introduce PathoPrompt, a framework that leverages Cross-Granular Semantic Alignment to improve sensitivity to refine the model's ability to capture subtle semantic variations in pathology image classification. Specifically, we introduce token-level fine-grained alignment, allowing the model to capture subtle differences that are crucial for accurate pathology image classification. Further, Cross-Granular Semantic Distillation improves the model's ability to generalize by filtering out irrelevant information from both coarse and fine-grained prompts. Moreover, PathoPrompt employs a prototype-based cross-modal separation mechanism, promoting distinct class boundaries by separating image and text semantics for more effective multi-modal representation learning. Experiments on five pathology datasets and three different task types demonstrate that our method achieves superior performance compared to previous methods.

1 Introduction

Recent advancements in pre-trained vision-language (V-L) models, such as CLIP[9] and ALIGN[3], have demonstrated impressive capabilities in generalizing across various downstream tasks. These models, which are trained on

[*] Co-corresponding author: wfsu@uic.edu.cn

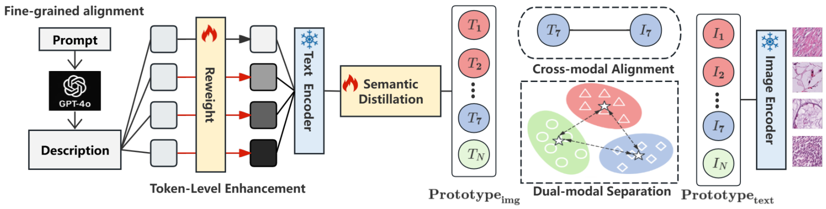

Fig. 1: **Comparison with previous method.** In contrast to prior work relying on sequence-level coarse-grained alignment, we propose token-level fine-grained alignment, enhancing sensitivity to subtle differences. To further improve generalization, we introduce Cross-Granular Semantic Distillation to filter irrelevant information. Additionally, a prototype-based cross-modal separation mechanism is established to enable effective dual-modal separation of image and text semantics, advancing multi-modal representation learning.

large-scale image-text pairs, can capture open-vocabulary concepts, offering significant flexibility. However, adapting these models to domain-specific tasks remains challenging due to their heavy reliance on carefully crafted input prompts. This dependence on manual prompt design not only limits efficiency but also fails to fully leverage the models' potential, especially in complex domains where more nuanced and detailed prompts are necessary for optimal performance.

Previous studies, such as HPT [11], have aimed to improve prompt learning by using hierarchical structures to illustrate relationships between visual elements (e.g., `"Water lily = flowers + leaves + blooms"`). However, these structured relationships do not exist in medical pathology images, where the complexity arises from subtle, non-hierarchical visual cues. Similarly, CoPrompt [10] investigated the use of large language models (LLMs) to generate descriptive prompts for aligning image and text representations. Nevertheless, these approaches depend on sequence-level, coarse-grained alignment (see **Figure 1**), which lacks the sensitivity required for fine-grained tasks such as medical pathology. In these instances, the inability to capture subtle variations often results in decreased performance.

In this work, we propose PathoPrompt, a framework designed to address the challenge of capturing subtle semantic differences that are crucial for accurate pathology image classification in medical V-L tasks. By implementing a token-level enhancement mechanism, our method enables more precise modeling of intra-class variations. Additionally, we introduce a Cross-Granular Semantic Distillation mechanism, which filters out irrelevant information from both coarse- and fine-grained prompts. This selective filtering process allows the model to focus on the most relevant features, enhancing its ability to generalize across unseen cases. Furthermore, we introduce a prototype-based dual-modal separation strategy to improve cross-modal alignment, enabling distinct and robust category representations in both image and text modalities. Our method enhances semantic separation across classes, which strengthens both modality alignment and class

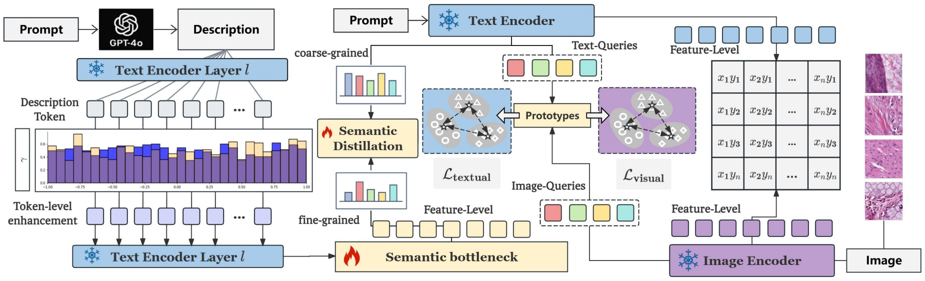

Fig. 2: **Overview of PathoPrompt structure.** (1) Token-level enhancement: Fine-grained descriptions undergo token-level adjustments, capturing subtle semantic details to increase sensitivity to nuanced inter-class variations. (2) Semantic Distillation: Coarse-grained prompts and fine-grained descriptions are distilled, aligning information from different levels of granularity. This dual-granularity distillation balances broad and detailed semantic information, improving robustness and precision. (3) Prototype-Based Separation: Prototypes are computed from features extracted by text and image encoders, facilitating inter-class separation and aligning representations across image and text modalities. This approach reinforces dual-modal alignment, supporting accurate cross-modal classification.

discrimination. This framework overcomes the limitations of existing methods, which often lack the sensitivity to capture subtle semantic differences essential for accurate pathology image classification, providing a more effective solution for fine-grained medical V-L tasks. The key contributions of our work are summarized below:(1) We propose PathoPrompt, a novel framework leveraging Cross-Granular Semantic Alignment to enhance sensitivity and refine the model's ability to capture subtle semantic variations in pathology image classification. (2) Token-level fine-grained alignment captures subtle differences, while Cross-Granular Semantic Distillation filters irrelevant information, improving generalization and sensitivity to inter-class variations. A prototype-based cross-modal separation mechanism further enhances inter-class discrimination by optimizing image-text semantic separation. (3) Experimental results on five pathology datasets across three task types demonstrate that our framework outperforms existing methods, achieving superior effectiveness and sensitivity for fine-grained medical tasks.

2 Method

2.1 Preliminaries

Let ϕ and θ denote the image and text encoders from CLIP, respectively. Given an input image $\boldsymbol{X} \in \mathbb{R}^{B\times H\times W}$, the input image $\boldsymbol{X}$ is divided into M fixed-size patches. Each patch is subsequently projected into a patch embedding $v_p \in \mathbb{R}^{M\times d_v}$ that represents the image in the latent space. The image encoder ϕ extracts its embedding as $\tilde{\boldsymbol{e}} = \phi(\boldsymbol{v_p})$. For a corresponding class label y, we generate a text prompt with a hand-crafted template, such as “a photo of a {CLASS}.” This template is tokenized as a sequence $\tilde{\boldsymbol{Y}} = \{\boldsymbol{t}_{SOS}, \boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{L-1}, \boldsymbol{t}_{EOS}\}$, where L is the token length, and $\boldsymbol{t}_{SOS}$ and $\boldsymbol{t}_{EOS}$ represent the learnable start and end token embeddings. The text encoder θ then processes $\tilde{\boldsymbol{Y}}$ through transformer layers to produce a class-specific text embedding $\tilde{\boldsymbol{g}} = \theta(\tilde{\boldsymbol{Y}})$, with $\tilde{\boldsymbol{g}} \in \mathbb{R}^{d_t}$. For inference, the image embedding $\tilde{\boldsymbol{e}}$ is compared against the class-specific text embeddings $\tilde{\boldsymbol{g}}$ for all C classes to find the most similar class. The probability of assigning class label y to image $\boldsymbol{X}$ is defined by:

$$p(y|\boldsymbol{X}) = \frac{\exp(\mathrm{sim}(\tilde{\boldsymbol{g}} \cdot \tilde{\boldsymbol{e}})/\tau)}{\sum_{i=1}^{C} \exp(\mathrm{sim}(\tilde{\boldsymbol{g}}_i \cdot \tilde{\boldsymbol{e}})/\tau)}. \tag{1}$$

where $\mathrm{sim}(\cdot)$ denotes cosine similarity, and τ is a temperature parameter.

2.2 Cross-Granular Semantic Distillation

In this section, we use Cross-Granular Semantic Distillation (GSD) to distill task-relevant information from two sources: prompts and descriptions. In the coarse-grained prompt, $\mathbf{t}_p \in \mathbb{R}^{C \times t}$, where each prompt follows the format "a photo of a [CLASS]" for each of the C classes. In fine-grained descriptions, $\mathbf{t}_d \in \mathbb{R}^{C \times s \times t}$, where each class has s detailed descriptions, such as " Debris is fragmented material scattered in the tissue.". To enhance robustness, we randomly select $\mathbf{t}_{d_{i=1}^{s}}$ from each batch for training. To distill task-relevant information from both coarse-grained prompts and fine-grained descriptions, we treat $\mathbf{t}_p$ and $\mathbf{t}_d$ as inputs with different granularities, where $\theta(\mathbf{t}_p)$ and $\mathbf{z}_d$ represent their corresponding embeddings. To enhance the model's sensitivity, we apply the token-level enhancement to the fine-grained descriptions $\mathbf{t}_d$. Specifically, each token embedding $\mathbf{t}_{\mathrm{d},i}$ is refined using token-specific parameters, $\tilde{\boldsymbol{\gamma}} = \{\tilde{\boldsymbol{\gamma}}_s, \tilde{\boldsymbol{\gamma}}_b\}$. These parameters enable fine-grained semantic adjustments for each token, which are crucial for achieving accurate cross-modal alignment. The enhanced description embedding is then computed as $\mathbf{z}_\mathrm{d} = \psi\, \theta(\tilde{\boldsymbol{\gamma}}^\top \mathbf{t}_\mathrm{d})$, where ψ represents the token-level semantic distillation mechanism. To align the coarse- and fine-grained representations, we introduce a loss function that ensures the learned embeddings are both compact and predictive of the target task. This objective is formulated as:

$$\mathcal{L}_\mathrm{desc} = D_\mathrm{KL}\left(P(y|\mathbf{z}_\mathrm{d}) \parallel P(y|\theta(\mathbf{t_p}))\right). \tag{2}$$

This loss function encourages the alignment of coarse-grained and fine-grained representations, ensuring that the learned embeddings are both compact and aligned with the target task, through token-level semantic distillation.

In contrast to prior work like CoPrompt, which employs a unidirectional alignment approach, our method preserves the domain-specific characteristics and knowledge embedded within the Description. Moreover, CoPrompt's adjustments operate at the sentence level, which limits the granularity of alignment. To address this, we employ a token-level alignment strategy, allowing for finer-grained, token-specific adjustments.

Our alignment loss $\mathcal{L}_{\text{align}}$ is defined as the cosine distance between the transformed and normalized Description and Prompt embeddings, formulated as:

$$\mathcal{L}_{\text{align}} = 1 - \frac{\psi\left(\theta\left(\sum_{i=1}^{L} \gamma_{s,i}\mathbf{t}_{\text{d},i} + \gamma_{b,i}\right)\right) \cdot \theta^{a}\left(\theta(\mathbf{t}_{\text{p}})\right)}{\left\|\psi\left(\theta\left(\sum_{i=1}^{L} \gamma_{s,i}\mathbf{t}_{\text{d},i} + \gamma_{b,i}\right)\right)\right\| \left\|\theta^{a}\left(\theta(\mathbf{t}_{\text{p}})\right)\right\|}. \tag{3}$$

Here, $\tilde{\boldsymbol{\gamma}}_s = \{\boldsymbol{\gamma}_s^{\text{SOS}}, \boldsymbol{\gamma}_s^{1}, \ldots, \boldsymbol{\gamma}_s^{L-1}, \boldsymbol{\gamma}_s^{\text{EOS}}\}$ and $\tilde{\boldsymbol{\gamma}}_b = \{\boldsymbol{\gamma}_b^{\text{SOS}}, \boldsymbol{\gamma}_b^{1}, \ldots, \boldsymbol{\gamma}_b^{L-1}, \boldsymbol{\gamma}_b^{\text{EOS}}\}$. θ^a is the adaptor layer of text encoder to transform the embedding vector. By minimizing this alignment loss, we ensure that both embeddings converge in the shared semantic space, capturing task-relevant details and maintaining domain-specific nuances. This approach enhances alignment robustness while avoiding the redundancy associated with sentence-level adjustments, thereby preserving and leveraging fine-grained semantic information in the description $\mathbf{t}_{\text{p}}$.

The final Cross-Granular Semantic Distillation GSD loss function combines the two terms, ensuring the model to capture domain-specific knowledge while preserving semantic precision. Therefore, the unified loss function can be expressed as follows:

$$\mathcal{L}_{\text{GSD}} = \mathcal{L}_{\text{desc}} + \mathcal{L}_{align}. \tag{4}$$

2.3 Prototype-Based Dual-Modal Separation

To achieve effective class separation across both image and text modalities, we propose a Prototype-Based Dual-Modal Separation mechanism. This method maintains embeddings for labeled samples, capturing both image and text features to compute class prototypes and optimize the semantic separation between classes.

To compute dual-modal prototypes, we calculate the mean feature for each class in both image and text modalities. For each class c, the class prototype $\mathbf{p}_c$ is computed as the mean of image embedding $\mathbf{v}_{\text{p}}$ and text embedding $\mathbf{t}_{\text{p}}$ corresponding to their class. Specifically,

$$\mathbf{p}_{\text{t},c} = \frac{1}{N_c}\sum_{i=1}^{N_c} \theta^{a}(\theta(\mathbf{v}_{\text{p,i}})), \quad \mathbf{p}_{\text{i},c} = \frac{1}{N_c}\sum_{i=1}^{N_c} \phi^{a}(\phi(\mathbf{t}_{\text{p,i}})). \tag{5}$$

The text feature $\mathbf{p}_{\text{text},c}$ for each class c is directly derived. Here, N_c denotes the number of samples belonging to class c, and the summation is performed over all indices i where $y_i = c$. The Prototype-Based Dual-Modal Separation (PDS) loss integrates inter-class separation across both modalities, enforcing dual-modal alignment to improve model generalization.

$$\mathcal{L}_{\text{PDS}} = \sum_{k=1}^{N_c-1}\sum_{j=k+1}^{N_c} \|\mathbf{p}_{t,k} - \mathbf{p}_{t,j}\|_2 + \sum_{k=1}^{N_c-1}\sum_{j=k+1}^{N_c} \|\mathbf{p}_{i,k} - \mathbf{p}_{i,j}\|_2. \tag{6}$$

Table 1: **Few-shot learning and Classes generalization.** Comparison between our method and SOTA methods for base-to-novel generalization on medical image classification datasets. Our method performs well over the compared methods. We use **red** and **blue** to indicate the first and second-best scores.

Tasks	**Kather**				**Colorectal**			
	Few-shot		Generalization		Few-shot		Generalization	
	Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
CoOp[13]	87.40	83.61	65.67	52.04	74.60	73.86	65.20	60.56
MaPLe[6]	85.42	80.63	77.83	68.47	76.70	75.94	70.30	62.12
PromptSRC[7]	88.08	83.68	79.18	75.30	78.10	77.57	74.80	70.90
CoPrompt[10]	85.24	79.25	76.80	70.45	74.40	72.63	66.80	60.18
PathoPrompt	89.86	85.27	83.26	76.48	78.20	77.71	70.60	66.86

Tasks	**BloodMNIST**				**KIMIA**			
	Few-shot		Generalization		Few-shot		Generalization	
	Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
CoOp[13]	68.75	64.23	54.87	52.07	78.52	74.56	60.74	53.25
MaPLe[6]	78.72	76.02	60.95	56.44	82.22	80.34	60.00	51.41
PromptSRC[7]	78.14	73.29	56.42	52.41	85.93	83.88	61.48	51.09
CoPrompt[10]	80.50	76.79	62.03	57.97	83.70	81.39	59.26	50.41
PathoPrompt	83.31	81.55	68.43	68.12	87.41	85.30	62.96	54.88

2.4 Final Loss

The cross-entropy loss $\mathcal{L}_{\text{CE}}$ is defined to measure the similarity between the normalized representations $\tilde{\boldsymbol{g}}$ and $\tilde{\boldsymbol{e}}$. This loss encourages higher similarity for the correct class while minimizing similarity with other classes, effectively enhancing class separation. It is defined as follows:

$$\mathcal{L}_{\text{CE}} = -\log \frac{\exp\left(\texttt{sim}\left(\tilde{\boldsymbol{g}} \cdot \tilde{\boldsymbol{e}}\right)/\tau\right)}{\sum_{j=1}^{C} \exp\left(\texttt{sim}\left(\tilde{\boldsymbol{g}} \cdot \tilde{\boldsymbol{e}}\right)/\tau\right)}. \tag{7}$$

The overall objective function combines multiple losses to achieve optimal model performance. The final loss is expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{GSD}} + \mathcal{L}_{\text{CE}} - \mathcal{L}_{\text{PDS}}. \tag{8}$$

3 Experiments

3.1 Experimental Setup

Few-shot Learning. To evaluate model performance under data-scarce conditions, we conducted few-shot experiments with a limited number of samples per class. **Classes Generalization.** To assess the model's capacity for generalizing across different classes, we divided the datasets into seen and unseen

Table 2: **Cross-Organ generalization.** Accuracy (%) evaluation for prompts learned from the source dataset. Our plugin consistently enhances existing prompt learning methods, whether textual, visual, or multi-modal.

	BD	BL	BR	CV	CO	ES	HN	KD	LV	LG	OV	PN	PT	SK	ST	TS	UT	AVG
CoOp [13]	70.00	86.49	78.44	58.33	58.65	75.44	66.67	52.78	80.36	61.90	87.18	86.00	64.58	60.00	78.95	89.36	81.48	72.74
MaPLe [6]	73.33	75.68	71.06	79.17	73.78	90.35	70.24	75.00	71.43	97.62	61.54	52.00	81.25	62.00	84.21	63.83	83.33	74.46
CoPrompt [10]	73.33	78.38	73.05	77.78	41.89	74.56	67.86	67.86	80.36	90.48	61.54	44.00	87.50	80.00	84.21	74.47	72.22	72.32
PromptSRC [7]	76.67	83.78	66.27	69.44	52.16	82.46	76.19	83.33	76.79	90.48	66.67	54.00	89.58	62.00	81.58	65.96	81.48	74.05
PathoPrompt	75.56	89.19	68.06	59.72	75.14	92.98	59.52	69.44	71.43	76.19	71.79	82.00	89.58	46.00	89.47	85.11	81.48	75.45

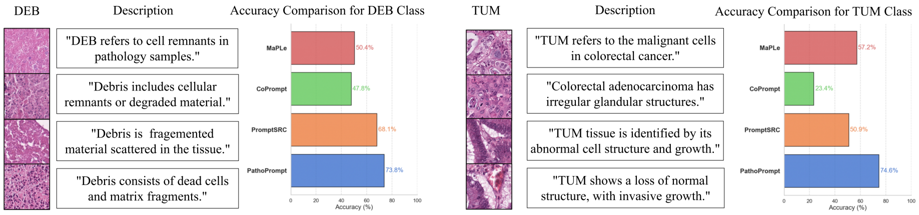

Fig. 3: **Accuracy comparison for DEB (debris) and TUM (tumor) classes** in the class generalization experiment on the Karther dataset. DEB is a seen class (trained with a few samples), while TUM is an unseen class (zero-shot).

categories. The model was trained in a few-shot setting using limited samples from the seen classes and evaluated on both seen (few-shot) and unseen (zero-shot) classes. **Cross-organ Generalization.** The model was trained with samples from adrenal gland tissues to classify benign and malignant cases and evaluated in zero-shot on unseen organs.

Datasets. The classes generalization and few-shot learning tasks utilized the Kather dataset [5], Colorectal Histology Dataset [4], BloodMNIST [12], and KIMIA Path960 [8], each contributing diverse histopathological images essential for assessing multi-class classification and texture-based learning. For the cross-organ generalization task, the PanNuke dataset [1] was used due to its comprehensive tissue type coverage and detailed annotations, enabling robust evaluation of cross-domain generalization.

Training Details. For a fair comparison, we adopt the CLIP-ViT-B/16 architecture as the base model across different methods. The model is initialized with pre-trained weights from Pathology Language and Image Pre-Training (PLIP) [2] to leverage domain-specific knowledge. We use the prompt template "a photo of a {class}" and train the model in fp16 precision. Training is conducted using the SGD optimizer with a learning rate of 0.0035 on a single GeForce RTX 4090.

3.2 Comparison to the State-of-the-Art (SOTA) Approaches

PathoPrompt consistently outperforms baseline models across multiple tasks, demonstrating superior performance in few-shot learning, class generalization, and cross-organ generalization. In the few-shot learning task, PathoPrompt achieves

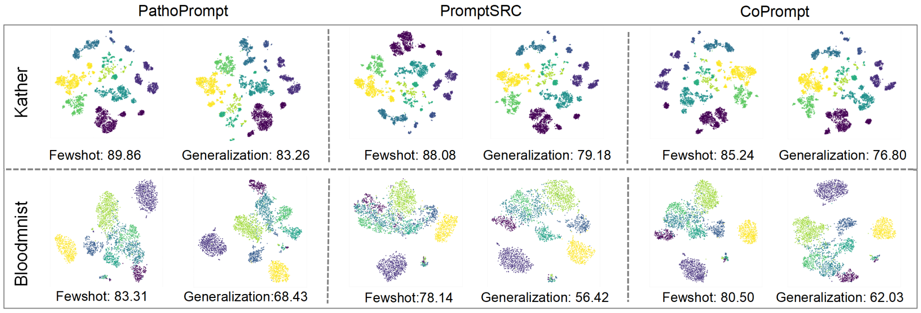

Fig. 4: **t-SNE visualizations of image embeddings** on the Kather and BloodMNIST datasets under few-shot and generalization settings.

Table 3: **Impact of Different Prompt Templates**.

Prompt Template	Accuracy
A photo of a {CLASS}	83.26
A tissue of {CLASS}	82.24
A microscopic view of {CLASS}	84.68
A tissue sample slide of {CLASS}	**85.60**

Table 4: **Ablation study results of components**.

Methods	Few.	Gen.	Cross.
Baseline	77.61	57.56	71.68
w/o GSD	82.20	60.77	66.35
w/o PDS	81.70	57.29	73.97
Ours	**83.31**	**68.43**	**75.45**

84.69% accuracy and 82.46% F1 score, surpassing PromptSRC by 2.13% in accuracy. For class generalization, it outperforms PromptSRC by 3.34% in accuracy, achieving 71.31% accuracy and 66.59% F1 score. In cross-organ generalization, PathoPrompt leads with 75.45% accuracy, surpassing the closest competitor, MaPLe, by 0.99%. Across all tasks, PathoPrompt excels in maintaining robust performance even with limited data, unseen classes, and cross-organ scenarios, proving its strong generalization ability for real-world medical applications, as seen in the accuracy comparison in Fig. 3 and t-SNE visualization in Fig. 4, which highlight PathoPrompt's superiority.

3.3 Ablation Study

Effect of Prompt Template Choice. Table 3 demonstrates that while a generic prompt on the Kather dataset, "`a photo of a [CLASS]`", achieves competitive performance, domain-specific prompts like "`a tissue sample slide of [CLASS]`" improve accuracy and F1 score, highlighting the advantage of fine-grained medical terminology in enhancing model performance for pathology classification.

Effect of Component. Table 4 shows few-shot generalization results on the BloodMNIST dataset and cross-organ generalization on PanNuke. Table 4 shows that PathoPrompt's full model outperforms ablated versions, with GSD and PDS both contributing significantly to few-shot accuracy, class generalization, and cross-organ generalization, demonstrating their crucial roles in enhancing generalization and inter-class separability.

4 Conclusion

We introduce PathoPrompt, a robust framework that enhances fine-grained pathology image classification by leveraging token-level alignment, Cross-Granular Semantic Distillation, and prototype-based cross-modal separation.

Acknowledgments. Our work was supported in part by the Guangdong Provincial Key Laboratory of IRADS (2022B1212010006) and in part by Guangdong Higher Education Upgrading Plan (2021-2025) and in part by by Guangdong and Hong Kong Universities "1+1+1" Joint Research Collaboration Scheme and in part by National Key Research and Development Program of China (2024YFF1206600) and the National Natural Science Foundation of China (U21A20520, 62325204).

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

References

1. Gamper, J., Koohbanani, N.A., Benes, K., Graham, S., Jahanifar, M., Khurram, S.A., Azam, A., Hewitt, K., Rajpoot, N.: Pannuke dataset extension, insights and baselines. arXiv preprint arXiv:2003.10778 (2020)
2. 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 pp. 1–10 (2023)
3. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. pp. 4904–4916. PMLR (2021)
4. Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Z"ollner, F.G.: Multi-class texture analysis in colorectal cancer histology. Scientific reports **6**, 27988 (2016)
5. Kather, J.N., Zöllner, F.G., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Weis, C.A.: Collection of textures in colorectal cancer histology. Zenodo https://doi. org/10 **5281** (2016)
6. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: CVPR. pp. 19113–19122 (2023)
7. Khattak, M.U., Wasim, S.T., Naseer, M., Khan, S., Yang, M.H., Khan, F.S.: Self-regulating prompts: Foundational model adaptation without forgetting. In: CVPR. pp. 15190–15200 (2023)
8. Kumar, M.D., Babaie, M., Zhu, S., Kalra, S., Tizhoosh, H.R.: A comparative study of cnn, bovw and lbp for classification of histopathological images. In: 2017 IEEE symposium series on computational intelligence (SSCI). pp. 1–7. IEEE (2017)
9. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)

10. Roy, S., Etemad, A.: Consistency-guided prompt learning for vision-language models. In: ICLR (2024), `https://openreview.net/forum?id=wsRXwlwx4w`
11. Wang, Y., Jiang, X., Cheng, D., Li, D., Zhao, C.: Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5749–5757 (2024)
12. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
13. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. arXiv preprint arXiv:2109.01134 (2021)