PathoTune: Adapting Visual Foundation Model to Pathological Specialists

Jiaxuan Lu[1], Fang Yan[1], Xiaofan Zhang[1,3], Yue Gao[2], and Shaoting Zhang[1,⋆]

[1] Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China
[2] Tsinghua University, Beijing 100084, China
[3] Shanghai Jiao Tong University, Shanghai 200240, China
{lujiaxuan,yanfang,zhangshaoting}@pjlab.org.cn

Abstract. As natural image understanding moves towards the pretrain-finetune era, research in pathology imaging is concurrently evolving. Despite the predominant focus on pretraining pathological foundation models, how to adapt foundation models to downstream tasks is little explored. For downstream adaptation, we propose the existence of two domain gaps, *i.e.*, the Foundation-Task Gap and the Task-Instance Gap. To mitigate these gaps, we introduce **PathoTune**, a framework designed to efficiently adapt pathological or even visual foundation models to pathology-specific tasks via multi-modal prompt tuning. The proposed framework leverages Task-specific Visual Prompts and Task-specific Textual Prompts to identify task-relevant features, along with Instance-specific Visual Prompts for encoding single pathological image features. Results across multiple datasets at both patch-level and WSI-level demonstrate its superior performance over single-modality prompt tuning approaches. Significantly, PathoTune facilitates the direct adaptation of natural visual foundation models to pathological tasks, drastically outperforming pathological foundation models with simple linear probing. The code is available at https://github.com/openmedlab/PathoDuet.

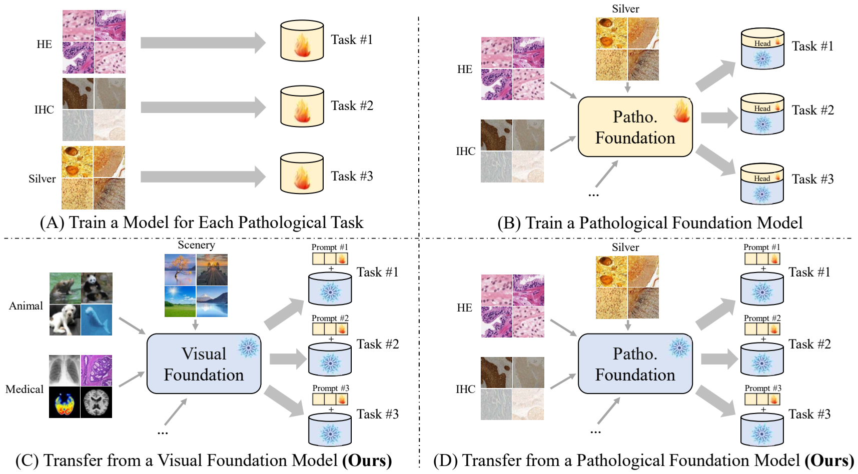
Keywords: Pathological Image · Prompt Tuning · Model Adaptation.

1 Introduction

Pathological image diagnostics stands as a critical step that informs clinical decisions by examining and interpreting stained images at the cellular level. Computational pathology integrates machine learning techniques that promise to revolutionize the approach to disease detection and analysis. In recent years, many deep learning-based pathology diagnostic methods have been explored, which can be categorized into patch-level [33, 32, 15] and WSI-level [30, 34, 22] frameworks. However, these models need to be individually trained for specific downstream tasks, *e.g.*, training a separate model to recognize breast cancer or Gleason grade of prostate, lacking in flexibility and generality.

⋆ S. Zhang is the corresponding author.

This work was supported by Shanghai Artificial Intelligence Laboratory.



(A) Train a Model for Each Pathological Task

(B) Train a Pathological Foundation Model

(C) Transfer from a Visual Foundation Model **(Ours)**

(D) Transfer from a Pathological Foundation Model **(Ours)**

Fig. 1. Compared to traditional paradigms of training separate models for each task or training a pathological foundation model, PathoTune directly adapts a visual or pathological foundation model to downstream tasks using multi-modal prompts.

As the field of language processing and image analysis has transitioned into the pretrain-finetune era, computational pathology has also entered the paradigm of foundation models and efficient finetuning, where it is desirable to train a generalized foundation model that can solve all downstream tasks. Recent advances have explored how self-supervised pretraining on large datasets can be used to develop pathological foundation models, including CTransPath [31], HIPT [4], Pathoduet [12], Virchow [29], *etc.* Despite these advancements, it is demonstrated that even with pathological foundation models, satisfactory performance cannot be achieved without finetuning [24]. While some works in the field of natural image analysis have explored the Parameter-Efficient Fine-Tuning (PEFT) [6, 20], how to effectively transfer the foundation models to downstream pathological tasks has received little attention.

This paper argues for the importance of efficiently adapting a generalist foundation model to downstream specialized models for pathological tasks. We propose that there exist two primary domain gaps in this process: the Foundation-Task Gap (FTG) and the Task-Instance Gap (TIG). FTG refers to the domain difference between the data encountered by the foundation model and the downstream pathological task or dataset, while TIG denotes the discrepancy between a specific image and the average distribution of images in the dataset, *e.g.*, varied staining variations inherent to each pathological image.

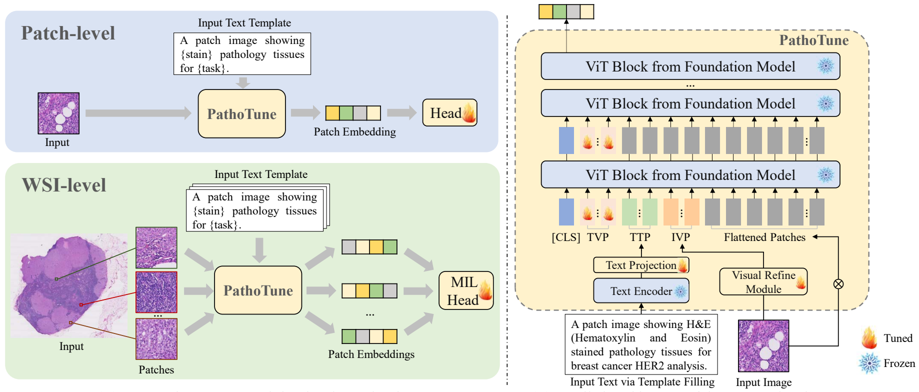
To address these challenges, we introduce **PathoTune**, a framework that employs multi-modal prompt tuning to adapt a foundation model for pathology-specific tasks with a minor parameter increment. The foundation model can be either a pretrained natural visual model or a pathological foundation model, as shown in Fig. 1. PathoTune leverages Task-specific Visual Prompts (TVP) and Task-specific Textual Prompts (TTP) to bridge the FTG by encoding task-related information. Additionally, it utilizes a Visual Refine Module to generate Instance-specific Visual Prompts (IVP) for addressing the TIG. The results from multiple datasets at both patch-level and WSI-level demonstrate that PathoTune not only outperforms state-of-the-art (SOTA) PEFT methods relying on single-modal prompts but also significantly exceeds elaborately pretrained pathological foundation models with linear probing.

2 Related Work

Traditional Pathology Modeling. In digital pathology, identifying cancer in Whole Slide Images (WSIs) poses a significant challenge due to their large size. Xu *et al.* [32] leverages CNNs pretrained on ImageNet to extract features from WSI patches for classification. To diagnose with WSI-level labels, Multi-Instance Learning (MIL) is utilized in a series of works [30, 34, 8], integrating CNNs with MIL for WSI classification. Additionally, Transformer models are being investigated for a more integrated WSI analysis by feeding features from numerous patches [23, 17]. Regardless of the backbone network, these methods require either training from scratch or full finetuning on a pretrained model, with a separate model needed for each specific task.

Pathological Foundation Model. With the emergence of foundation models in natural language processing [28, 3], computer vision [11, 10], *etc.*, recent studies have explored the development of pathological foundation models based on self-supervised learning. Studies like Huang *et al.* [13] and Ciga *et al.* [7] apply contrastive learning to pathological patches. CTransPath [31] enhances the MoCo v3 framework with a pseudo positive selection mechanism to improve similarity handling between patches. Pathoduet [12] builds on MoCo v3 with additional pretraining tasks for cross-scale and cross-stain challenges. Large-scale data utilization includes HIPT's [4] hierarchical pyramid ViT pretraining on 10,678 WSI slides, UNI [5] employing 100,000 slides with the DINO v2 framework, and Virchow [29] using 1.5 million slides.

Parameter-Efficient Fine-Tuning. PEFT has become a prominent and efficient alternative in natural language processing [1, 20], offering accuracy comparable to full finetuning but with fewer parameters and reduced storage. In computer vision, Adaptformer [6] finetunes visual adapters in foundation models for diverse tasks. Prompt tuning [14, 27] emerges as an alternative, enabling task transfer without altering the network's structure. Jia *et al.* introduces VPT [14] with learnable tokens as visual prompts, and Sohn *et al.* [26] proposes generating prompt tokens using a generator. VQT [27] utilizes "query"-only learnable tokens for further parameter reduction. How to adapt the foundation model to pathological downstream tasks is worth exploring as well.


Patch-level
Input Text Template
A patch image showing {stain} pathology tissues for {task}.
Input
PathoTune
Patch Embedding
Head
WSI-level
Input Text Template
A patch image showing {stain} pathology tissues for {task}.
Input
Patches
PathoTune
Patch Embeddings
MIL Head
PathoTune
ViT Block from Foundation Model
ViT Block from Foundation Model
ViT Block from Foundation Model
[CLS]
TVP
TTP
IVP
Flattened Patches
Text Projection
Text Encoder
Visual Refine Module
A patch image showing H&E (Hematoxylin and Eosin) stained pathology tissues for breast cancer HER2 analysis.
Input Text via Template Filling
Input Image
Tuned
Frozen

(A) Input and Output at Patch-level and WSI-level

(B) Detailed Architecture of the Proposed PathoTune

Fig. 2. Overview of the proposed PathoTune. (A) The input and output of PathoTune for both patch-level and WSI-level tasks. (B) Detailed architecture of PathoTune, encompassing the Task-specific Visual Prompts (TVP), Task-specific Textual Prompts (TTP) and Instance-specific Visual Prompts (IVP).

3 Methodology

3.1 Problem Formulation

Assuming that the data distribution of the natural image and the pathology image is represented as F and D, respectively, the corresponding visual foundation model and the ideal pathology model are represented as $\Phi(\cdot)$ and $\Psi(\cdot)$. In the adaptation of foundation models to downstream tasks, we identify the existence of two domain gaps: the Foundation-Task Gap and the Task-Instance Gap.

- Foundation-Task Gap (FTG): The gap between the data domain F pretrained by the foundation model $\Phi(\cdot)$ and downstream pathological domain D, which is relevant to the specific task.
- Task-Instance Gap (TIG): Domain gap between each instance image in the task-specific dataset and the dataset's average data distribution, including nuances such as staining and glandular structure variations.

In downstream adaptation, the FTG not only encompasses the discrepancy between the natural image domain F and the pathology domain D, but also reflects the significant divergence between the visual foundation model $\Phi(F)$ and the ideal pathology model $\Psi(D)$. To bridge this gap, we propose the employment of task-specific prompts P_{task}, which are designed to adapt the visual foundation model $\Phi(F)$ to the pathology domain D, denoted as $\Phi(D; P_{task})$. The proposed task-specific prompts P_{task} are aimed at minimizing

$$\min_{P} \|\Phi(D; P_{task}) - \Psi(D)\| \,. \tag{1}$$

Particularly, the $\Phi(\cdot)$ can be either a visual or a pathological foundation model.

For each individual instance image $x \in D$ within a pathological dataset, the TIG is quantified as the variance $\sigma(D)$, representing the dispersion of the dataset's distribution. To depict the specificity of the embedding of an instance image $\Phi(x; P_{task})$ compared to the mean value $\Phi(\overline{x}; P_{task})$, we introduce the instance-specific prompts P_{ins} which are designed to minimize

$$\min_{P} \mathbb{E}_{x \in D} \left[\| \Phi(x; P_{ins}) - \Psi(x) \| \right], \tag{2}$$

where $\Psi(x)$ denotes the embedding derived from feeding image x into the ideal pathology model.

3.2 Multi-Modal Prompt Design

In response to the two domain gaps inherent in downstream pathological adaptation, the proposed PathoTune introduces three kinds of prompts, including Task-specific Visual Prompts (TVP), Task-specific Textual Prompts (TTP), and Instance-specific Visual Prompts (IVP). The TVP and TTP are designed as task-specific prompts, with the purpose of relieving the Foundation-Task Gap. Conversely, the IVP serves as the instance-specific prompts. From a modal perspective, both TVP and IVP are categorized as visual prompts, while TTP operates as textual prompts. The complete pipeline as well as the inputs and outputs of PathoTune are shown in Fig. 2.

Task-specific Visual Prompts. To mitigate the Foundation-Task Gap, it is crucial to convey task-specific information to the foundation model. In this context, we interpret the visual prompt explored in existing works [14, 27] as a type of "soft" prompt with learnable task-specific information relevant to the downstream pathological domain. Specifically, the Task-specific Visual Prompts (TVP) introduces several learnable tokens into each layer of the Vision Transformer (ViT). For the ViT with L layers, let $P^{l}_{TVP} \in \mathbb{R}^{N \times C}$ be the matrix of learnable tokens at layer l, where N is the number of tokens and C is the token dimension. The corresponding N TVP tokens P^{l}_{TVP} are prepended to the patch embedding E^{l} before being fed into the l-th layer.

Task-specific Textual Prompts. In addition to the "soft" visual prompts that promote token self-learning, we consider textual descriptions as another approach profiling the downstream pathological task and dataset. The proposed Task-specific Textual Prompts (TTP) $P_{TTP} \in \mathbb{R}^{T \times C}$ utilizes a text template filled with specific stain (*e.g.*, HE or IHC) and task information to generate text embeddings, which are then aligned with other tokens through a frozen text encoder θ^{TE} and a tunable text projection layer θ^{TP}. Assuming the text template be P_{text} = "A patch image showing {stain} pathology tissues for {task}". The text embedding for a given task is obtained as

$$P_{TTP} = f_{TP} \left(f_{TE}(P_{text}; \theta_{TE}); \theta_{TP} \right), \tag{3}$$

where θ_{TE} and θ_{TP} are the parameters of the text encoder and the text projection layer, respectively. The text projection layer not only formalizes the feature dimension, but also aligns the text features with flattened patches and prompts.

Instance-specific Visual Prompts. The Instance-specific Visual Prompts (IVP) $P_{IVP} \in \mathbb{R}^{M \times C}$ targets the Task-Instance Gap (TIG) by capturing unique characteristics of individual pathological instances. Specifically, we propose a lightweight Visual Refine Module (VRM) f_{VRM} to extract the specific staining and glandular features relative to a single patch image. For a given instance image $x \in D$, the VRM processes it into a coarse-grained embedding, which is then replicated into M tokens, expressed as

$$P_{IVP} = f_{VRM}(x; \theta_{VRM}), \tag{4}$$

where θ_{VRM} denotes the tunable parameters within the VRM. Different from the TVP and TTP which have fixed tokens for a specific dataset, the IVP is instance-wise, with tokens generated for each input image.

3.3 Overall Procedure

The proposed PathoTune appends the above three types of prompts to the input of the ViT structure derived from the foundation model. Assuming that the pathological image is flattened into K tokens, the embedding of the l-th layer is expressed as $E^{l-1} \in \mathbb{R}^{K \times C}$. Thus, the first layer of ViT can be represented as

$$[V^1, E^1] = \Phi^1 \left([V^0, P^0_{TVP}, P_{TTP}, P_{IVP}, E^0] \right), \tag{5}$$

where V^l is the [CLS] token, and $\Phi^l(\cdot)$ denotes the Transformer layer of the l-th layer. Unlike the first layer which requires three types of tokens, subsequent layers only need to replace the TVP tokens using P^{l-1}_{TVP}, expressed as

$$[V^l, E^l] = \Phi^l \left([V^{l-1}, P^{l-1}_{TVP}, E^{l-1}] \right), \tag{6}$$

where the last layer of the [CLS] token V^L is fed into the tunable patch-level or WSI-level head to classify pathological images. With the proposed prompts, we can reuse the extensive knowledge embedded in the foundation model, requiring only finetuning the prompts and the head for adaptation with a far lesser number of trainable parameters. Additionally, compared to previous paradigms (Fig. 1 (A) and (B)) required to train a specialized pathology model $\Psi(\cdot)$ based on the pathological dataset D, the proposed paradigm offers the support for a multitude of tasks through a shared foundation model augmented by specialized prompts.

4 Experiments and Results

Datasets. We conduct a comprehensive evaluation of PathoTune across extensive pathology datasets, including both public datasets, *i.e.*, BCI [19], NCT [16], SICAPv2 [25], and the private RJ-Prost dataset. These datasets span patch-level (BCI, NCT) and WSI-level (SICAPv2, RJ-Prost) tasks, covering various organs and staining types (HE and IHC). Among them, RJ-Prost is a proprietary dataset from an anonymous hospital focusing on prostate Gleason grading,

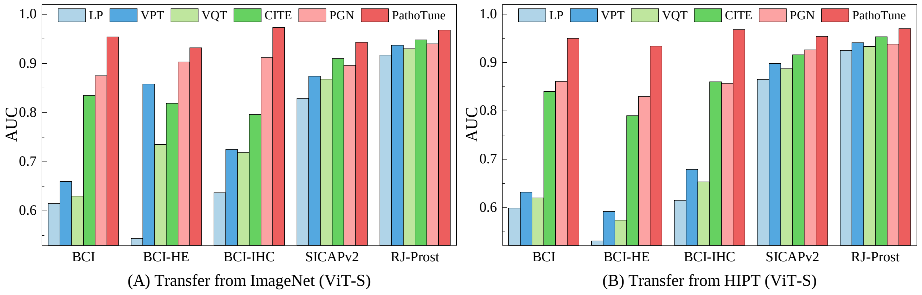
Table 1. Ablation results (%) based on different foundation models on multiple datasets, where "FT" stands for full finetuning and "LP" stands for linear probing.

Found.	Mode	Prompts			Patch-level						WSI-level			
					BCI-HE		BCI-IHC		NCT		SICAPv2		RJ-Prost	
		TTP	TVP	IVP	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
ImageNet (ViT-S)	FT				94.2	76.0	97.0	85.4	99.8	95.3	94.2	74.9	96.8	77.2
	LP				54.4	15.7	63.7	26.3	97.4	84.5	82.9	11.8	91.7	51.1
	Ours	✓			63.1	16.2	68.0	30.1	98.5	89.9	84.5	25.0	93.1	55.6
			✓		85.8	67.5	72.5	31.1	99.2	91.5	87.4	59.0	93.7	60.9
				✓	92.5	75.5	96.0	81.5	99.4	92.1	91.3	68.3	95.2	74.0
		✓	✓	✓	**93.2**	**76.1**	**97.3**	**84.3**	**99.7**	**92.4**	**94.3**	**74.8**	**96.8**	**76.4**
HIPT (ViT-S)	FT				94.7	79.9	95.3	82.1	99.8	94.2	95.5	79.9	97.3	80.7
	LP				53.1	14.5	61.5	29.0	98.4	84.9	86.5	46.9	92.5	54.7
	Ours	✓			58.2	15.8	65.5	31.0	98.9	90.2	87.3	55.3	93.5	58.8
			✓		59.2	15.7	67.9	32.2	99.1	91.3	89.8	64.9	94.1	67.4
				✓	92.8	72.7	90.1	65.1	99.6	92.4	92.9	72.5	96.3	74.8
		✓	✓	✓	**93.4**	**75.4**	**96.8**	**82.8**	**99.8**	**94.0**	**95.4**	**79.3**	**97.0**	**80.5**
ImageNet (ViT-B)	FT				95.1	81.4	97.5	86.4	99.7	95.2	97.2	83.5	97.5	82.9
	LP				64.2	18.5	69.0	36.5	98.6	89.8	95.3	73.5	93.9	56.8
	Ours	✓			65.1	20.0	71.2	33.4	98.9	90.0	95.8	81.0	94.8	65.9
			✓		66.5	21.2	75.9	32.9	99.4	90.6	96.7	82.3	95.2	70.3
				✓	93.5	77.0	97.0	83.7	99.7	91.9	97.0	82.5	96.1	78.3
		✓	✓	✓	**94.0**	**77.6**	**97.3**	**84.6**	**99.8**	**92.0**	**97.5**	**84.2**	**97.2**	**82.4**
Pathoduet (ViT-B)	FT				98.7	89.7	99.0	92.5	99.7	94.9	97.3	84.5	97.8	83.0
	LP				68.0	27.4	75.1	38.7	99.3	93.9	94.1	70.2	94.6	69.0
	Ours	✓			71.0	30.4	78.4	40.1	99.4	94.0	94.8	78.3	95.1	72.1
			✓		75.5	36.2	82.0	46.2	99.4	94.2	95.2	80.2	95.8	72.3
				✓	92.6	76.8	96.6	83.6	99.6	94.0	96.7	82.8	96.5	80.2
		✓	✓	✓	**94.1**	**77.6**	**97.3**	**86.9**	**99.8**	**94.9**	**97.6**	**84.8**	**97.5**	**82.6**

which includes 1,042 WSIs and four categories: negative, grade 3, grade 4, and grade 5. The BCI dataset which contains both HE and IHC stains is divided into "BCI-HE" and "BCI-IHC" for specific experiments. For dataset division, BCI adheres to the official guideline, with identical validation and test sets. NCT follows the protocol established by Bian *et al* [2]. For remaining datasets without standardized division protocols, we allocate data into training, validation, and test sets in a 7:2:1 ratio. 4-fold cross-validation is applied to all except BCI.

Implementations. We transfer both visual foundation models (ViT pretrained on ImageNet) and pathological foundation models (HIPT [4] and Pathoduet [12]) to each downstream dataset individually. HIPT employs ViT-S, while Pathoduet utilizes ViT-B as the backbones for comparisons. The text encoder in TTP leverages the pretrained BERT [9], while the VRM module in TVP initializes with the first 4 layers of ResNet18. In most experiments, the token number for TVP, TTP, and IVP is set at 10, 2, and 2, respectively, with a batch size of 32. We employ the RAdam [18] optimizer at a learning rate of 0.0002.

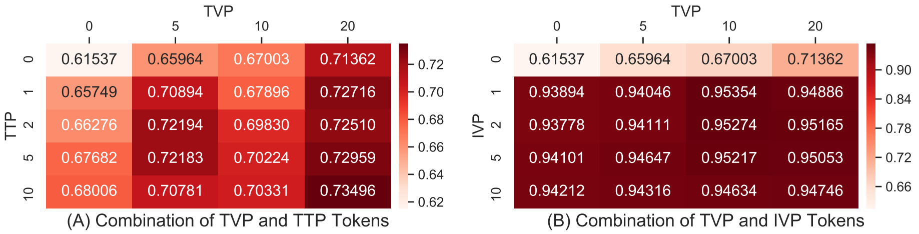
Effectiveness of PathoTune. The results of different foundation models using mixed combinations of prompts (Table 1) yield key insights: (1) Our method significantly surpasses linear probing in all scenarios, closely rivals full finetuning with just 5.9% of the trainable parameters (More details in supplementary material). Furthermore, employing multi-modal prompts greatly enhances performance compared to single-modal usage. (2) The performance of



(A) Transfer from ImageNet (ViT-S)

(B) Transfer from HIPT (ViT-S)

Fig. 3. Comparisons of the PathoTune with other SOTA methods of PEFT.



(A) Combination of TVP and TTP Tokens

(B) Combination of TVP and IVP Tokens

Fig. 4. Comparisons of the PathoTune with different prompt combination.

a useful downstream adaptation method (*e.g.*, our PathoTune) based on the natural visual foundation model far exceeds that of a poor downstream adaptation approach (*e.g.*, linear probing) based on pathological foundation model, suggesting that efficient downstream adaptation is even more important than pretraining a pathological foundation model. (3) Transferring from a pathological foundation model shows slightly better results than a visual foundation model under the same backbone scale, indicating the optimal strategy originates from a pathological foundation model paired with an effective finetuning approach. (4) PathoTune can significantly enhance underperforming foundation models, *e.g.*, ImageNet (ViT-S), elevating their performance to rival that of specialized models, *e.g.*, Pathoduet (ViT-B).

Comparisons with SOTAs. We evaluate PathoTune's performance against SOTA methods including VPT [14], VQT [27], CITE [35], and PGN [21] across various datasets, as depicted in Fig. 3. PathoTune consistently surpasses all compared methods, regardless of whether it transfers from the visual or pathological foundation model. The results demonstrate the superiority of PathoTune's multi-modal prompts elaborated for domain gaps over these approaches using single-modal prompts.

Impacts of prompt combination. We evaluate the performance of PathoTune under different combinations of prompts adapted from ImageNet (ViT-S) on the BCI dataset, with IVP and TTP taken as 0 respectively. As shown in Fig. 4, using a combination of prompts yields better results than using a single prompt, and IVP emerges as the most effective one.

5 Conclusion

In this paper, we present **PathoTune**, an innovative framework designed to adapt generalist foundation models to specialized pathological tasks through multi-modal prompt tuning. By addressing the Foundation-Task Gap and the Task-Instance Gap, we propose the Task-specific Visual Prompts, Task-specific Textual Prompts, and Instance-specific Visual Prompts. PathoTune not only surpasses SOTA methods but also remarkably outperforms pretrained pathological foundation models using linear probing, providing a new paradigm for computational pathology applications in the pretrain-finetune era.

Acknowledgments. This study was supported by Shanghai Artificial Intelligence Laboratory.

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

References

1. Ben Zaken, E., Goldberg, Y., Ravfogel, S.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In: Annual Meeting of the Association for Computational Linguistics. pp. 1–9 (2022)
2. Bian, H., Shao, Z., Chen, Y., Wang, Y., Wang, H., Zhang, J., Zhang, Y.: Multiple instance learning with mixed supervision in gleason grading. In: MICCAI. pp. 204–213 (2022)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS **33**, 1877–1901 (2020)
4. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: CVPR. pp. 16144–16155 (2022)
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: A general-purpose self-supervised model for computational pathology. arXiv preprint arXiv:2308.15474 (2023)
6. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. NeurIPS **35**, 16664–16678 (2022)
7. Ciga, O., Xu, T., Martel, A.L.: Self supervised contrastive learning for digital histopathology. Machine Learning with Applications **7**, 100198 (2022)
8. Courtiol, P., Tramel, E.W., Sanselme, M., Wainrib, G.: Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. arXiv preprint arXiv:1802.02212 (2018)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)

11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
12. Hua, S., Yan, F., Shen, T., Zhang, X.: Pathoduet: Foundation models for pathological slide analysis of h&e and ihc stains. arXiv preprint arXiv:2312.09894 (2023)
13. Huang, Z., Chai, H., Wang, R., Wang, H., Yang, Y., Wu, H.: Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In: MICCAI. pp. 561–570 (2021)
14. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: ECCV. pp. 709–727 (2022)
15. Källén, H., Molin, J., Heyden, A., Lundström, C., Åström, K.: Towards grading gleason score using generically trained deep convolutional neural networks. In: International Symposium on Biomedical Imaging. pp. 1163–1167 (2016)
16. Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo10 **5281** (2018)
17. Li, H., Yang, F., Zhao, Y., Xing, X., Zhang, J., Gao, M., Huang, J., Wang, L., Yao, J.: Dt-mil: deformable transformer for multi-instance learning on histopathological image. In: MICCAI. pp. 206–216 (2021)
18. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265 (2019)
19. Liu, S., Zhu, C., Xu, F., Jia, X., Shi, Z., Jin, M.: Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In: CVPR. pp. 1815–1824 (2022)
20. Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., Tang, J.: P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In: Annual Meeting of the Association for Computational Linguistics. pp. 61–68 (2022)
21. Loedeman, J., Stol, M.C., Han, T., Asano, Y.M.: Prompt generation networks for efficient adaptation of frozen vision transformers. arXiv preprint arXiv:2210.06466 (2022)
22. Pal, S., Valkanas, A., Regol, F., Coates, M.: Bag graph: Multiple instance learning using bayesian graph neural networks. In: AAAI. vol. 36, pp. 7922–7930 (2022)
23. 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. NeurIPS **34**, 2136–2147 (2021)
24. Sikaroudi, M., Hosseini, M., Gonzalez, R., Rahnamayan, S., Tizhoosh, H.: Generalization of vision pre-trained models for histopathology. Scientific Reports **13**(1), 6065 (2023)
25. Silva-Rodríguez, J., Colomer, A., Sales, M.A., Molina, R., Naranjo, V.: Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. Computer Methods and Programs in Biomedicine **195**, 105637 (2020)
26. Sohn, K., Chang, H., Lezama, J., Polania, L., Zhang, H., Hao, Y., Essa, I., Jiang, L.: Visual prompt tuning for generative transfer learning. In: CVPR. pp. 19840–19851 (2023)
27. Tu, C.H., Mai, Z., Chao, W.L.: Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In: CVPR. pp. 7725–7735 (2023)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
29. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Mathieu, P., van Eck, A., Lee, D., Viret, J., et al.: Virchow: A million-slide digital pathology foundation model. arXiv preprint arXiv:2309.07778 (2023)

30. Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Tsougenis, E., Huang, Q., Cai, M., Heng, P.A.: Weakly supervised deep learning for whole slide lung cancer image analysis. Transactions on Cybernetics **50**(9), 3950–3962 (2019)
31. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis **81**, 102559 (2022)
32. Xu, Y., Jia, Z., Ai, Y., Zhang, F., Lai, M., Eric, I., Chang, C.: Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation. In: International Conference on Acoustics, Speech and Signal Processing. pp. 947–951 (2015)
33. Xu, Y., Jia, Z., Wang, L.B., Ai, Y., Zhang, F., Lai, M., Chang, E.I.C.: Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics **18**, 1–17 (2017)
34. Xu, Y., Li, Y., Shen, Z., Wu, Z., Gao, T., Fan, Y., Lai, M., Chang, E.I.C.: Parallel multiple instance learning for extremely large histopathology image analysis. BMC Bioinformatics **18**, 1–15 (2017)
35. Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D.: Text-guided foundation model adaptation for pathological image classification. In: MICCAI. pp. 272–282 (2023)