FEW-SHOT WHOLE SLIDE PATHOLOGY CLASSIFICATION WITH MULTI-GRANULAR VISION-LANGUAGE MODELS

Anh-Tien Nguyen[1], **Duy H. M. Nguyen**[2,3,4*], **Nghiem T. Diep**[2 *], **Trung Q. Nguyen**[2], **Nhat Ho**[5], **Jacqueline M. Metsch**[6], **Miriam C. Maurer**[6], **Daniel Sonntag**[2,7], **Hanibal Bohnenberger**[6], **Anne-Christin Hauschild**[1,8]

[1]Justus-Liebig-Universität Gießen
[2]German Research Center for Artificial Intelligence (DFKI)
[3]University of Stuttgart
[4]Max Planck Research School for Intelligent Systems (IMPRS-IS)
[5]The University of Texas at Austin
[6]University Medical Center Göttingen
[7]Oldenburg University
[8]Institut für Predictive Deep Learning for Medicine and Healthcare
[*] Co-second contribution
{anne-christin.hauschild}@uni-giessen.de

ABSTRACT

In this study, we propose a novel architecture for a large vision-language model adapted with *a multi-granular prompt learning method* to advance few-shot pathology classification. Starting with the Prov-GigaPath foundation model - pre-trained on 1.3 billion pathology image patches - *we extend it into a vision-language model by adding adaptors* and aligning it with medical text encoders via contrastive learning on 923K image-text pairs. In contrast to previous approaches that combine prompts with frozen features using prefix embeddings or self-attention, *our multi-granular attention mechanism* evaluates interactions between learnable prompts, individual image patches, and patch groups, capturing both fine details and broader context. We further improve the precision with an *unbalanced optimal transport-based visual-text distance* that mitigates perturbations from data augmentation. Experiments on lung and kidney pathology imaging modalities show that our method outperforms state-of-the-art competitors and improves performance across various architectures, including CLIP, PLIP, and the Prov-GigaPath integrated PLIP. We provide pre-trained weights at this link.

1 INTRODUCTION

Whole slide images (WSIs) provide high-resolution views of tissue samples and are the gold standard for cancer diagnostics and treatment. However, they can contain billions of pixels, making annotation and interpretation costly. Obtaining access to sufficiently large, annotated datasets is particularly challenging for cancer subtypes, which has spurred the development of few-shot and weakly supervised methods (Madabhushi & Lee, 2016; Li et al., 2023; Lin et al., 2023; Ryu et al., 2023; Shi et al., 2024) and

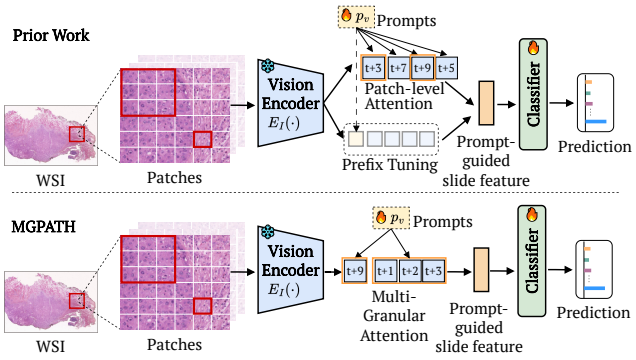

Figure 1: Unlike previous methods that add prompts at prefix positions or patch-level attention - disrupting structural correlations - our MGPATH framework integrates prompts at both regional and individual patch levels (multi-granular attention).

especially multiple-instance learning
(MIL) (Ilse et al., 2018; Xu et al., 2019; Li et al., 2023; Lin et al., 2023; Tang et al., 2023; Shi et al., 2023). Yet, MIL can struggle to capture meaningful regions in complex tissue structures.

Visual language models (VLMs) (Lu et al., 2023; Huang et al., 2023; Ikezogwo et al., 2024) address this by integrating slide-level features and textual descriptions. Specifically, VLMs utilize multi-scale images (Shi et al., 2024; Han et al., 2024), enabling the extraction of both global and local features from whole slide images (WSIs) at various resolutions. To efficiently adapt a pre-trained vision-language model, prompt learning (Zhou et al., 2022b; Gao et al., 2024) is applied, where learnable prompts are incorporated into the input text to guide the model. Additionally, contextual prompts (Li & Liang, 2021; Yao et al., 2024) are embedded into feature representations using a self-attention mechanism (Vaswani, 2017).

Although effective, those approaches face some challenges. **First**, (i) prompt learning with frozen visual features often overlooks the hierarchical interactions between prompts and visual features, particularly multi-granular attention between prompts and both individual patches and patch groups. This limitation weakens the model's ability to capture dependencies across scales, from fine-grained details to broader context, reducing its effectiveness in understanding complex pathology patterns. **Second**, (ii) many VLMs are currently based on `CLIP` (Radford et al., 2021), which lacks explicit pre-training on pathology images, restricting adaptability in few-shot settings, especially when prompt learning is applied with a frozen architecture. While `PLIP` (Huang et al., 2023), trained on 200k pathology image-text pairs or `CONCH` (Lu et al., 2024) (1.17M), has shown improvements, it remains unclear whether scaling to larger pathology-specific samples would yield further gains. **Lastly**, (iii) most whole-slide pathology VLMs rely on cosine similarity for vision-text alignment, which struggles with multiple text descriptions for sub-regions (Chen et al., 2023) and data perturbations (Nguyen et al., 2024b), limiting its ability to capture fine-grained alignments.

We present MGPATH, a vision-language model (VLM), namely MGPATH, designed for whole-slide pathology classification. Building on `Prov-GigaPath` (Xu et al., 2024), which is *pre-trained on 1.3 billion pathology patches*, we *extend it into a VLM* using contrastive learning. This is achieved by integrating the `PLIP` text encoder (Huang et al., 2023), trained on 200K pathology image-text pairs, through a *parameter-efficient adaptor*. We strengthen alignment using 923K additional image-text pairs from `ARCH` (Gamper & Rajpoot, 2021), `PatchGastricADC22` (Tsuneki & Kanavati, 2022), and `Quilt-1M` (Ikezogwo et al., 2024), training only lightweight adaptors. Next, we introduce multi-granular prompt learning for few-shot WSI tasks by generating visual embeddings and descriptive text prompts for image patches at multiple resolutions using large language models (Han et al., 2024; Shi et al., 2024; Qu et al., 2024). Unlike prior methods that concatenate patches or apply conventional attention (Li & Liang, 2021; Zhou et al., 2022b; Yao et al., 2024; Shi et al., 2024), our approach integrates learnable prompts with frozen visual features at both fine- and coarse-grained levels (Figure 1). We model WSI patches as a spatial graph, using bounding box coordinates for region-level aggregation via message passing. This structure is encoded as Key-Value tokens, which interact with Query embeddings from prompts. By directing attention across both patch and region levels, our method effectively captures hierarchical information, enhancing feature representation and refining focus on critical tissue areas.

Finally, we leverage the optimal transport (OT) (Nguyen et al., 2021; Pham et al., 2020; Séjourné et al., 2023; Chen et al., 2023; Dong et al., 2023; Nguyen et al., 2024b; Zhan et al., 2021) to define the distance between prompt-fused visual embeddings and multiple text prompts, which flexibly aligns heterogeneous data distributions. This is particularly useful for few-shot WSI classification, as OT (i) adapts to data augmentation and noise while preserving structural relationships and (ii) handles modality imbalances, especially when text prompts describe only sub-regions of WSI samples. Across three datasets and multiple architectures (`CLIP-ResNet50`, `CLIP-ViTB16`, `PLIP`, and `(Prov-GigaPath)-PLIP`), MGPATH consistently outperforms 14 state-of-the-art MIL and VLM models. Notably, MGPATH (Prov-GigaPath-PLIP) surpasses `MSCPT` (Han et al., 2024) by 5% in F1 and 8% in AUC on `TCGA-BRCA` and outperforms foundation VLMs `CONCH` (Lu et al., 2024) and `QUILT` (Ikezogwo et al., 2024) by 6% in accuracy on the same dataset.

2 RELATED WORKS

Large-scale Pre-trained Models for Pathology. Recent advances in large-scale pre-trained pathology models fall into two categories: vision models and VLMs. Vision models like `Virchow` (Ike-

zogwo et al., 2024), Hibou (Nechaev et al., 2024), UNI (Chen et al., 2024), and Prov-GigaPath (Xu et al., 2024) learn robust visual features from massive datasets, with Prov-GigaPath (1.3B patches) excelling in high-resolution tissue analysis. VLMs such as PLIP (Huang et al., 2023) (200K image-text pairs), CONCH (Lu et al., 2024) (1.17M), and QUILTNET (Ikezogwo et al., 2024) (1M) combine images and text for enhanced pathology interpretation. Our MGPATH bridges these approaches by using a parameter-efficient adaptor to integrate Prov-GigaPath (the largest vision encoder) with a VLM text encoder (e.g., PLIP or CONCH), leveraging both rich visual and semantic features. While our experiments use PLIP for consistency with baselines, the method is adaptable to larger pre-trained text models.

Few-shot learning in WSI. MIL models WSIs as bags of patches, aggregating features via non-parametric methods like mean/max pooling. However, these can diminish critical disease-related signals. To improve relevance, attention-based MIL, GNNs, and Transformers have been explored (Lu et al., 2021; Chen et al., 2021; Ilse et al., 2018; Li et al., 2021; Shao et al., 2021; Zheng et al., 2022). Meanwhile, VLMs use contrastive learning to align image-text pairs, enhancing pathology tasks despite the challenge of collecting large-scale labeled data. Models like MI-Zero, PLIP, and CONCH have been trained on hundreds of thousands to over a million pathology image-text pairs (Lu et al., 2023; Huang et al., 2023; Lu et al., 2024). Some also integrate multi-magnification images and multi-scale text to mimic pathologists' workflows (Shi et al., 2024; Han et al., 2024). Our MGPATH builds on VLMs, amplifying the advantages of large pre-trained pathology models while introducing *parameter-efficient multi-granular prompt learning* to improve few-shot adaptation.

Prompt Learning for Vision-Language Adaptation. Prompt tuning is key to adapting large pre-trained models, as seen in multimodal systems like CLIP. Instead of handcrafted templates, methods such as CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), and MaPLe (Khattak et al., 2023) learn prompts for domain generalization (Ge et al., 2023; Yao et al., 2024), knowledge prototypes (Zhang et al., 2022b; Li et al., 2024), or diversity (Lu et al., 2022; Shu et al., 2022). However, these focus on natural images rather than the multi-scale, structurally complex data in WSIs. While some methods (Shi et al., 2024; Qu et al., 2024) apply prompts via self-attention to frozen features, they can miss intricate tissue structures. By contrast, our multi-granular prompt learning framework applies *attention to both individual patches and spatial groups*, better aligning with WSIs' hierarchical complexity.

3 METHOD

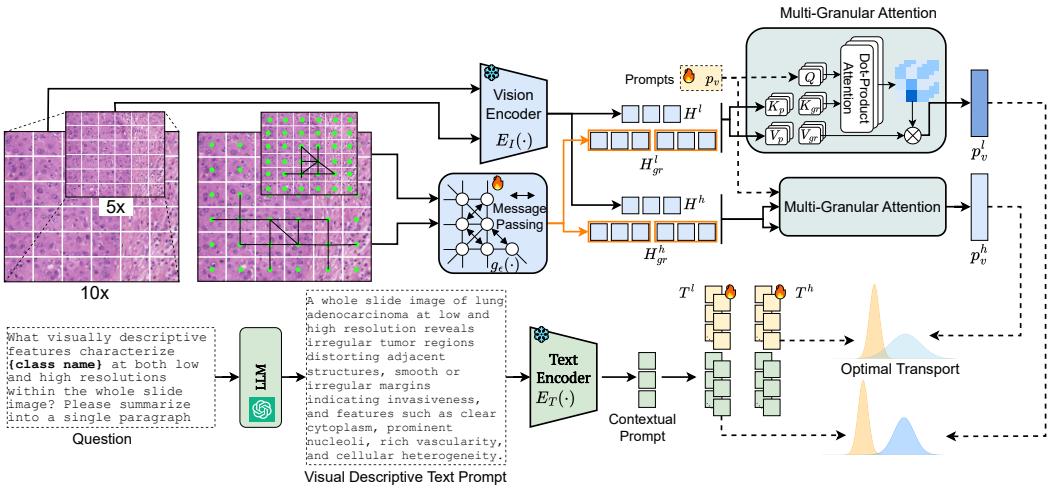

Figure 2: The pipeline of the proposed MGPATH method. Low- and high-resolution image patches are processed with large language models to generate contextual descriptions (Section 3.2). Visual prompts are integrated with frozen features through multi-granular attention at both patch and group-of-patch levels 3.3. The final output is obtained by aligning visual and text embeddings using optimal transport (Section 3.4).

3.1 BRIDGING PATHOLOGY VISUAL AND TEXT ENCODERS

To leverage `Prov-GigaPath`'s pre-trained visual features, we implement lightweight adaptors that map patch-level image features into the embedding space of the `PLIP` text encoder. These adaptors enable joint image-text training with minimal parameter updates, since only the adaptor weights are fine-tuned.

Given pathology image-text pairs $\{(\mathbf{I}_i, \mathbf{T}_i) | \, i = 1, 2.., N\}$, let $E_I(.)$ be the `Prov-GigaPath` vision encoder for patch-level features, and $E_T(.)$ the `PLIP` text encoder. For each batch of size B, the image and text embeddings are $\mathbf{x}_i = E_I(\mathbf{I}_i) \in \mathbb{R}^{d_v}$, $\mathbf{t}_i = E_T(\mathbf{T}_i) \in \mathbb{R}^{d_t}$.

We then design two trainable adaptors, $A_I(.)$ and $A_T(.)$, to project $(\mathbf{x}_i, \mathbf{t}_i)$ into a shared dimension $\mathbb{R}^d$, optimizing the noise contrastive loss (Oord et al., 2018):

$$\mathcal{L}_{con} = \mathbb{E}_B \left[-\log \frac{\exp\left(\cos(A_I(\mathbf{x}_i), A_T(\mathbf{t}_i))/\tau\right)}{\sum_j \exp\left(\cos(A_I(\mathbf{x}_i), A_T(\mathbf{t}_j))/\tau\right)} \right], \tag{1}$$

where $\cos(.)$ is the cosine similarity, and τ denotes for temperature of the softmax function. Both the `Prov-GigaPath` vision encoder and the `PLIP` text encoder remain frozen, while only $A_I(.)$ and $A_T(.)$ are trained. Once Eq. equation 1 is optimized, the adaptor outputs serve as visual and text embeddings for downstream tasks. We refer to this model as `GigaPath-PLIP`.

3.2 MULTI-MAGNIFICATION DESCRIPTIVE TEXT PROMPTS

Designing effective text prompts is crucial for enhancing vision-language models (VLMs) in whole-slide image (WSI) analysis. Pathologists typically assess WSIs by first examining tissue structures at low magnification, then zooming in to observe finer details such as nuclear shape and size. Recent works (Shi et al., 2024; Han et al., 2024) have harnessed this multi-scale approach by introducing dual-scale descriptive text prompts and leveraging large language models (LLMs), yielding considerable gains in classification performance. Building on this idea, we further refine and extend the strategy to boost model effectiveness. The prompt template is described in Figure 3 where {`class name`} is replaced by specific categories.

LLM Prompt

```
What visually descriptive features characterize {class name} at
both low and high resolutions within the whole-slide image?  Please
summarize into a single paragraph.
```

Figure 3: LLM template prompt.

Next, at each low/high scale, rather than inserting a single learnable text prompt of length K alongside a frozen contextual prompt obtained from LLMs (Shi et al., 2024; Han et al., 2024), we propose using M learnable prompts. This strategy captures different sub-regions or structural features within each patch that might be missed by a single prompt. Specifically, we define visual descriptive text prompts for both low- and high-resolution scales as follows:

$$\begin{aligned} \mathbf{T}_i^{(l)} &= \left\{ \left([\omega_i^{(l)}]_1 \, [\omega_i^{(l)}]_2 \, ... [\omega_i^{(l)}]_K [\text{LLM context}] \right) |_{i=1}^{M} \right\} \\ \mathbf{T}_i^{(h)} &= \left\{ \left([\omega_i^{(h)}]_{(1)} \, [\omega_i^{(h)}]_2 \, ... [\omega_i^{(h)}]_K [\text{LLM context}] \right) |_{i=1}^{M} \right\}, \end{aligned} \tag{2}$$

where $[\omega_i^{\beta}]_j, j \in [1, ..., K], i \in [1, .., M]$ are KM *trainable textual prompts* for each resolution $\beta \in \{l, h\}$.

3.3 GRANULARITY-AWARE VISUAL PROMPT LEARNING

For each WSI W, we denote by $\left\{W^{(l)}, W^{(h)}\right\}$ are representations of W at low and high magnification. We define a bag of multiple instances of W as $I = \left\{I^{(l)}, I^{(h)}\right\}$ where $I^{(l)} \in \mathbb{R}^{N_l \times N_b \times N_b \times 3}$, $I^{(h)} \in \mathbb{R}^{N_h \times N_b \times N_b \times 3}$ with N_l, N_h indicate the number of low and high-resolution image patches

and N_b is the patch size. We use a non-overlapping sliding window technique to extract patches I. from the WSI.

3.3.1 PATCHES-BASED PROMPTING

A frozen image encoder $E_I(.)$ (or $A_I(E_I(.))$ for `GigaPath-PLIP`) maps each patch I into feature vectors $H = \{H^{(l)} \in \mathbb{R}^{N_l \times d}, H^{(h)} \in \mathbb{R}^{N_h \times d}\}$ where d is the feature dimension. To consolidate the large set of patch features into a final slide-level representation, we introduce a set of learnable visual prompts $\mathbf{p}_v \in \mathbb{R}^{N_p \times d}$, which progressively merge patch features in $H^{(l)}$. Concretely, we treat $\mathbf{p}_v$ as the QUERY and all features in $H^{(l)}$ as the KEYS $K_p^{(l)}$ and VALUES $V_p^{(l)}$ in a self-attention mechanism Vaswani (2017). We then associate $\mathbf{p}_v$ with the patch features as:

$$\mathbf{p}_{v,p}^{(l)} = \text{Normalize}\left(\text{SoftMax}\left(\frac{\mathbf{p}_v {K_p^{(l)}}^T}{\sqrt{d}}\right) V_p^{(l)}\right) + \mathbf{p}_v\,. \tag{3}$$

3.3.2 SPATIAL PATCH GROUP-BASED PROMPTING

To quantify spatial correlations across multiple instances of I, we extract the coordinates for all its patches. Let $I^{(l)} = \left\{I_1^{(l)}, I_2^{(l)}, ..., I_{N_l}^{(l)}\right\}$ denote the patches and $H^{(l)} = \left\{H_1^{(l)}, H_2^{(l)}, ..., H_{N_l}^{(l)}\right\}$ their corresponding features. We construct a graph $G^{(l)} = (V^{(l)}, E^{(l)})$ to capture regional tissue structure, where $V^{(l)} = I^{(l)}$, and $E^{(l)} \in \{0,1\}^{N_l \times N_l}$. Edges in $E^{(l)}$ are defined by linking each path to its $K-$ nearest neighbors in the coordinate space. We set the node feature embedding $X^{(l)} = H^{(l)} \in \mathbb{R}^{N_l \times d}$, so each vertex $v_i^{(l)}$ is associated with a feature $x_i^{(l)} = H_i^{(l)}$.

We design a trainable message-passing network $g_\epsilon(.)$ using the graph attention layer (GAT) Veličković et al. (2017) to capture the feature representation of each node and its local neighbors. The GAT layer performs message passing as follows:

$$\begin{aligned}
\alpha_{i,j} &= \frac{\exp\left(\sigma(a_s^T \Theta_s\, x_i^{(l)} + a_t^T \Theta_t\, x_j^{(l)})\right)}{\sum_{k \in \mathcal{N}(i)\cup\{i\}} \exp(\sigma(a_s^T \Theta_s\, x_i^{(l)} + a_t^T \Theta_t\, x_k^{(l)}))} \\
x_i^{(l)'} &= \alpha_{i,i} \Theta_s\, x_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j} \Theta_t\, x_j^{(l)},
\end{aligned} \tag{4}$$

where $x_i^{(l)'}$ is aggregated features of $x_i^{(l)}$ with its local region after GAT layer, $\sigma(.)$ is the `LeakyReLU` activation function, $\mathcal{N}(i)$ denote the neighboring nodes of the i-th node, $\alpha_{i,j}$ are the attention coefficients and $a_s, a_t, \Theta_s, \Theta_t$ are weight parameters of $g_\epsilon(.)$.

After performing message passing with $g_\epsilon(.)$, we obtain an updated graph $G^{(l)'}$, where each node encapsulates its respective local feature region. We then aggregate all the feature nodes in $G^{(l)'}$ into a single vector $H_{gr}^{(l)}$, which acts as another set of KEYS $K_{gr}^{(l)}$ and VALUES $V_{gr}^{(l)}$ for region-level features. Following the same approach as equation 3, we associate the prompt $\mathbf{p}_v$ with these group-level features:

$$\mathbf{p}_{v,gr}^{l} = \text{Normalize}\left(\text{SoftMax}\left(\frac{\mathbf{p}_v {K_{gr}^{(l)}}^T}{\sqrt{d}}\right) V_{gr}^{(l)}\right) + \mathbf{p}_v. \tag{5}$$

The final output of our multi-granular is computed as:

$$\mathbf{p}_v^{(l)} = (1-\alpha) \cdot \mathbf{p}_{v,p}^{(l)} + \alpha \cdot \mathbf{p}_{v,gr}^{(l)}\,. \tag{6}$$

3.4 OPTIMAL TRANSPORT FOR VISUAL-TEXT ALIGNMENT

In this study, we employ OT to measure the alignment between visual prompt-guided slide features $\mathbf{p}_v^{(l)}$ and $\mathbf{p}_v^{(h)}$, and descriptive text prompts $\mathbf{T}^{(l)}$ and $\mathbf{T}^{(h)}$. Although OT has been explored for prompt learning in natural images and multi-modal learning (Kim et al., 2023; Chen et al., 2023;

Nguyen et al., 2024a; Séjourné et al., 2023), we are the first to adapt it for whole-slide imaging (WSI), effectively handling the alignment of multi-magnification patches to capture rich structural details across scales.

Recap OT. Given two sets of points (features), we can represent the corresponding discrete distributions as follows:

$$\boldsymbol{\mu} = \sum_{i=1}^{M} p_i \delta_{f_i}, \quad \boldsymbol{\nu} = \sum_{j=1}^{N} q_j \delta_{g_i}, \tag{7}$$

where δ_f and δ_g represent Dirac delta functions centered at $\boldsymbol{f}$ and $\boldsymbol{g}$, respectively, and M and N indicate the dimensions of the empirical distribution. The weight vectors $\boldsymbol{p} = \{p_i\}_{i=1}^{M}$ and $\boldsymbol{q} = \{q_i\}_{j=1}^{N}$ lie within the M and N-dimensional simplex, respectively, meaning they satisfy $\sum_{i=1}^{M} p_i = 1$ and $\sum_{j=1}^{N} q_j = 1$. The discrete optimal transport problem can then be expressed as:

$$\begin{aligned} \boldsymbol{T}^* = & \underset{\boldsymbol{T} \in \mathbb{R}^{MXN}}{\arg\min} \sum_{i=1}^{M} \sum_{j=1}^{N} \boldsymbol{T}_{ij} \boldsymbol{C}_{ij} \\ \text{s.t.} \quad & \boldsymbol{T}\mathbf{1}^N = \boldsymbol{\mu}, \quad \boldsymbol{T}^\top \mathbf{1}^M = \boldsymbol{\nu}. \end{aligned} \tag{8}$$

where $\boldsymbol{T}^*$ is denoted as the optimal transport plan, which is optimized to minimize the total distance between the two probability vectors, $\boldsymbol{C}$ is the cost matrix which measures the distance between $\boldsymbol{f}_i$ and $\boldsymbol{g}_j$. We then define the OT distance between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ as:

$$d_{\mathrm{OT}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \langle \boldsymbol{T}^*, \boldsymbol{C} \rangle. \tag{9}$$

Objective functions. Given the visual prompt-guided slide features $\mathbf{p}_v^{(l)} \in \mathbb{R}^{N_p \times d}$ in equation 6 and the text prompts $\mathbf{T}^{(l)}$ in equation 2, we obtain the textual embedding $\mathbf{p}_t^{(l)}$ by applying E_T to $\mathbf{T}^{(l)}$, i.e., $\mathbf{p}_t^{(l)} = E_T(\mathbf{T}^{(l)})$. Let $\mathbf{T}_c^{(l)}$ denote the input text prompts for class c, $\left(\mathbf{p}_t^{(l)}\right)_c$ be the corresponding textual embedding, and $\left(\mathbf{p}_v^{(l)}\right)_c$ be the visual prompt-guided slide features associated with the same class c. We apply OT to minimize the distance between $\mathbf{T}_c^{(l)}$ and $\left(\mathbf{p}_v^{(l)}\right)_c$, denoted by $d_{\mathrm{OT}}\left(\mathbf{T}_c^{(l)}, \left(\mathbf{p}_v^{(l)}\right)_c\right)$. Then, the cost matrix $\boldsymbol{C}$ is computed as $\boldsymbol{C} = \left(\mathbf{1} - \boldsymbol{F}^T \boldsymbol{G}\right) \in \mathbb{R}^{M \times N_p}$, where $\left(\mathbf{p}_t^{(l)}\right)_c \rightarrow \boldsymbol{F} = \left\{\boldsymbol{f}_i|_{i=1}^{M}\right\}$ and $\left(\mathbf{p}_v^{(l)}\right)_c \rightarrow \boldsymbol{G} = \left\{\boldsymbol{g}_j|_{j=1}^{N_p}\right\}$. We can produce $d_{\mathrm{OT}}\left(\mathbf{T}_c^{(h)}, \left(\mathbf{p}_v^{(h)}\right)_c\right)$ by using the same procedure at high-resolution image patches. Then, the prediction probability is written as:

$$\mathrm{P}_c = \frac{\exp(2 - \sum_{k \in \{l,h\}} d_{\mathrm{OT}}\left(\mathbf{T}_c^{(k)}, \left(\mathbf{p}_v^{(k)}\right)_c\right))}{\sum_{c'=1}^{C} \exp(2 - \sum_{k \in \{l,h\}} d_{\mathrm{OT}}\left(\mathbf{T}_{c'}^{(k)}, \left(\mathbf{p}_v^{(k)}\right)_c\right))}, \tag{10}$$

where λ_k controls contribution of each-resolution. Finally, we can train the model with the cross-entropy as:

$$\mathcal{L}_{class} = \mathrm{Cross}(\mathrm{P}, \mathrm{GT}), \tag{11}$$

with $\mathrm{Cross}(.)$ be the cross-entropy and GT denotes slide-level ground-truth.

4 EXPERIMENTS

Datasets for contrastive learning. `PatchGastricADC22` Tsuneki & Kanavati (2022) contains about 262K patch-level images from H&E-stained gastric adenocarcinoma specimens, each paired with diagnostic captions from Mita Hospital, Japan. `QUILT-1M` Ikezogwo et al. (2024) comprises approximately 653K images and one million pathology image-text pairs obtained from 1,087 hours of educational histopathology videos on YouTube. `ARCH` Gamper & Rajpoot (2021) provides a multiple-instance captioning dataset featuring bag- and tile-level pathology images. For our contrastive training, we focus on tile-level samples from these datasets, yielding roughly 923K total images.

Downstream tasks. We evaluated our method on two TCGA datasets: `TCGA-NSCLC` and `TCGA-RCC` obtained from the Cancer Genome Atlas Data Portal The Cancer Genome Atlas (TCGA). We follow the data splitting settings of `ViLa-MIL` Shi et al. (2024) for dividing `TCGA-NSCLC` and `TCGA-RCC` into training, validation, and test sets.

4.1 RESULTS

Comparison to State-of-the-Art. We compare our MGPATH with state-of-the-art multi-instance learning methods, including `Maxpooling`, `Mean-pooling`, `ABMIL` (Ilse et al., 2018), `CLAM` (Lu et al., 2021), `TransMIL` (Shao et al., 2021), `DSMIL` (Li et al., 2021), `GTMIL` (Zheng et al., 2022), `DTMIL` (Zhang et al., 2022a), `RRT-MIL` (Tang et al., 2024) and `IBMIL` (Lin et al., 2023), and vision-language methods, including `CoOp` (Zhou et al., 2022b), `CoCoOp` (Zhou et al., 2022a), `Metaprompt` (Zhao et al., 2024), `TOP` (Qu et al., 2024), `ViLa-MIL` (Shi et al., 2024), `MSCPT` (Han et al., 2024), `QUILT` (Ikezogwo et al., 2024), `CONCH` (Lu et al., 2024). Among these, `QUILT` and `CONCH` are foundation VLMs.

MGPath versions. We offer several versions of our MGPATH, including the `CLIP` backbone with `ResNet-50` (`CLIP50`) for `TCGA-NSCLC` and `TCGA-RCC`. Additionally, we provide a version using the `PLIP` backbone, as well as our proposed `GigaPath-PLIP` pre-trained models.

Table 1: Comparison of methods on TCGA-NSCLC and TCGA-RCC datasets with few-shot settings. Results are shown for AUC, F1, and Accuracy (ACC).

Methods	# Param.	TCGA-NSCLC			TCGA-RCC		
		AUC	F1	ACC	AUC	F1	ACC
Max-pooling	197K	53.0±6.0	45.8±8.9	53.3±3.4	67.4±4.9	46.7±11.6	54.1±4.8
Mean-pooling	197K	67.4±7.2	61.1±5.5	61.9±5.5	83.3±6.0	60.9±8.5	62.3±7.4
ABMIL Ilse et al. (2018)	461K	60.5±15.9	56.8±11.8	61.2±6.1	83.6±3.1	64.4±4.2	65.7±4.7
CLAM-SB Lu et al. (2021)	660K	66.7±13.6	59.9±13.8	64.0±7.7	90.1±2.2	75.3±7.4	77.6±7.0
CLAM-MB Lu et al. (2021)	660K	68.8±12.5	60.3±11.1	63.0±9.3	90.9±4.1	76.2±4.4	78.6±4.9
TransMIL Shao et al. (2021)	2.54M	64.2±8.5	57.5±6.4	59.7±5.4	89.4±5.6	73.0±7.8	75.3±7.2
DSMIL Li et al. (2021)	462K	67.9±8.0	61.0±7.0	61.3±7.0	87.6±4.5	71.5±6.6	72.8±6.4
GTMIL Zheng et al. (2022)	N/A	66.0±15.3	61.1±12.3	63.8±9.9	81.1±13.3	71.1±15.7	76.1±12.9
DTMIL Zhang et al. (2022a)	986.7K	67.5±10.3	57.3±11.3	66.6±7.5	90.0±4.6	74.4±5.3	76.8±5.2
IBMIL Lin et al. (2023)	N/A	69.2±7.4	57.4±8.3	66.9±6.5	90.5±4.1	75.1±5.2	77.2±4.2
ViLa-MIL Shi et al. (2024)	8.8M/47M	74.7±3.5	67.0±4.9	67.7±4.4	92.6±3.0	78.3±6.9	80.3±6.2
CONCH (Lu et al. (2024))	110M	89.46±10.2	78.5±9.31	78.78±9.1	88.08±4.59	78.21±4.2	71.67±19.4
QUILT Ikezogwo et al. (2024)	63M	79.66±13.19	72.30±13.35	72.42±13.24	96.92±1.6	78.46±5.55	86.34±1.56
MGPATH (CLIP)	1.6M/39M	77.2±1.3	70.9±2.0	71.0±2.1	92.1 ± 2.8	76.5 ± 5.2	81.7 ± 2.9
MGPATH (PLIP)	592K	83.6 ± 4.5	76.41 ± 4.8	76.5 ± 4.8	94.7 ± 1.6	78.6 ± 4.9	83.6 ± 3.5
MGPATH (PLIP-G)	5.35M	**93.02±2.99**	**84.64±4.75**	**84.77±4.67**	**98.2±0.31**	**88.33±3.41**	**91.72±1.74**

Table 2: Ablation studies on multi-granular (M-Gran), ratio combines two attention levels (α in Eq (6)).

Configurations	TCGA-NSCLC		
	AUC	F1	ACC
MGPATH (CLIP)	76.2±2.2	69.0±3.5	69.3±2.8
- w/o M-Gran (CLIP)	74.6±2.2	67.8±2.4	67.8±2.5
MGPATH (PLIP-G)	91.7±3.6	84.2±4.6	84.4±4.5
- w/o M-Gran (PLIP-G)	90.6±4.5	82.4±5.7	82.5±5.7
MGPATH, $\alpha = 0.2$	76.2±2.2	69.0±3.5	69.3±2.8
- $\alpha = 0.5$	73.7±3.1	67.4±2.6	67.8±2.7
- $\alpha = 0.8$	72.2±5.2	66.4±5.5	66.8±5.2
	TCGA-RCC		
MGPATH (CLIP)	92.1±2.8	76.5±5.2	81.7±2.9
- w/o M-Gran (CLIP)	91.6±3.5	72.3±6.4	80.2±4.4
MGPATH (PLIP-G)	98.1±0.6	85.7±1.1	89.9±2.0
- w/o M-Gran (PLIP-G)	98.1±0.6	85.0±4.0	89.3±3.0

Observed Results on Few-shot and Zero-shot Settings. As shown in Table 1, MGPATH, based on `CLIP50`, achieves top recording performances and providing significant improvements over other VLMs with similar architectures such as `ViLa-MIL`. Furthermore, `PLIP` backbone particularly improved MGPATH. For instance, on TCGA-NSCLC using backbone `CLIP50`, MGPATH achieves an accuracy of 71.0%, compared to 67.0% of `ViLa-MIL`. Moreover, using the `PLIP` backbone provides an additional 6% improvement on TCGA-NSCLC, demonstrating MGPATH's adaptability and effectiveness across different backbones.

By incorporating distilled pathology features from `Prov-GigaPath` Xu et al. (2024) — pre-trained on 1.3 billion pathology images — MGPATH(PLIP-G) achieves new state-of-the-art accuracies of 84.77% on TCGA-NSCLC and 91.72% on TCGA-RCC.

MGPATH also establishes a new benchmark in *zero-shot tasks*, demonstrating its ability to generalize without additional fine-tuning. As shown in Table 5, it achieves the highest average performance across two datasets, outperforming state-of-the-art foundation VLMs, including `CONCH` and `PLIP`.

4.2 ABLATION STUDIES

PLIP enhanced Prov-GigaPath. We evaluate the performance of our proposed `PLIP-G` under the following settings. (i) using vision-language PLIP model; (ii) using our pre-trained architectures through adaptors; (iii) `Prov-GigaPath` integrated with `PLIP` through randomly initialized adaptor layers, (iv) `Prov-GigaPath` with an adaptor layer mapping to class output, training only the MLP and last FFN layer. Table 3 demonstrates that combining `Prov-GigaPath` with `PLIP` by pre-trained adaptors (Section 3.1) improves performance over using either model individually.

Multi-Granular Prompt Learning. In Table 2, MGPATH with multi-granular (M-Gran) outperforms the variant without it (rows 1–2 for `CLIP` and 3–4 for `PLIP-G`) on `TCGA-NSCLC`, with a similar trend observed on `TCGA-RCC`. The table also indicates that a 0.2/0.8 ratio of graph-based to prototype-guided attention yields the best performances.

OT as Alignment between Contextual Prompts. Table 4 confirms the benefits of incorporating OT into MGPATH on the TCGA-NSCLC and TCGA-RCC datasets. Notably, using OT (rows 1 and 2) boosts performance compared to cosine similarity (rows 3 and 4). Moreover, the results indicate that the optimal number of prompt vectors can vary by dataset.

Table 3: Ablation studies on adaptor learning for Prov-GiGaPath and PLIP . PLIP-G denotes for mixed version between Prov-GiGaPath and PLIP.

Methods	# Param.	TCGA-NSCLC		
		AUC	F1	ACC
MGPATH (PLIP)	592K	83.6±4.5	76.41±4.8	76.5±4.8
MGPATH (PLIP-G)	5.35M	91.7±3.6	84.2±4.6	84.4±4.5
MGPATH Random Adaptors	5.35M	91.4±4.2	82.8±5.7	83.0±5.6
GIGAPATH Tuning (MLP + last FFN)	4.7M	62.7±3.5	64.66±5.3	52.8±3.4

Table 4: Contribution of OT and multiple descriptive text prompts

Methods	TCGA-NSCLC		
	AUC	F1	ACC
MGPATH (OT, 4 text prompts)	76.2±2.2	69.0±3.5	69.3±2.8
MGPATH (OT, 2 text prompts)	77.2±1.3	70.9±2.0	71.0±2.1
MGPATH (Cosine, 2 text prompts)	75.8±3.7	68.3±4.5	68.4±4.5
	TCGA-RCC		
MGPATH (OT, 4 text prompts)	92.1±2.8	76.5±5.2	81.7±2.9
MGPATH (OT, 2 text prompts)	92.1±2.6	75.6±3.9	80.4±2.4
MGPATH (Cosine, 4 text prompts)	91.8±2.8	75.9±4.3	80.5±2.6

Table 5: Zero-shot classification performance on TCGA-NSCLC, and TCGA-RCC. Metrics include balanced accuracy (B-Acc) and weighted F1-score (W-F1).

Zero-shot	TCGA-NSCLC		TCGA-RCC		Average	
	B-Acc	W-F1	B-Acc	W-F1	B-Acc	W-F1
QuiltNet	61.3	56.1	59.1	51.8	57.23	49.33
CONCH	**80.0**	**79.8**	72.9	69.1	72.3	70.03
PLIP	70.0	68.5	50.7	46.0	61.8	59.43
PLIP-G (Our)	72.7	72.6	**81.3**	**81.4**	**74.67**	**74.63**

4.3 DISCUSSION

In this study, we propose MGPATH, which achieves significant improvements in few-shot and zero-shot WSI classification across multiple datasets. However, we have not explored other potential challenges, leaving room for further investigation in future work. For instance, integrating VLMs models with other pathology foundation models such as CONCH or extending the approach to segmentation tasks.

5 CONCLUSION

Whole slide images have become indispensable in clinical practice — particularly for cancer diagnosis — analyzing their complex, hierarchical, high-resolution structures remains a significant challenge for automated methods. Although recent VLMs research leveraging few-shot and weakly supervised learning has achieved promising results with limited annotations, these approaches often overlook the hierarchical relationships among the learnable prompts, individual patches, and patch groups. Furthermore, they lack the precision needed to capture fine-grained alignments between image-text pairs. In this study, we propose MGPATH, a VLM that integrates Prov-GigaPath with PLIP, to overcome these limitations. Our granular prompt learning approach effectively captures hierarchical tissue interactions, resulting in significant improvements in WSI classification. Experimental results demonstrate that our MGPATH achieves state-of-the-art results in WSIs classification. We expect this work to inspire future research on integrating vision-language models with multi-granular prompt learning to capture local, global, and spatial information in WSI structures while leveraging optimal transport methods to enhance few-shot learning in pathology.

6 ACKNOWLEDGEMENT

The study was supported by the *German Ministry of Education and Research (BMBF)* under grant agreement *No. 01KD2208A (project FAIrPaCT)*.

The authors gratefully acknowledge the computing time granted by the Resource Allocation Board and provided on the supercomputer Emmy/Grete at NHR-Nord@Göttingen as part of the NHR infrastructure. The calculations for this research were conducted with computing resources under the project *nim00014*.

The authors gratefully acknowledge the computing time granted by the KISSKI project. The calculations for this research were conducted with computing resources under the project *kisski-umg-fairpact-2*. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Duy M. H. Nguyen. Duy M. H. Nguyen and Daniel Sonntag are also supported by the XAINES project (BMBF, 01IW20005), No-IDLE project (BMBF, 01IW23002), and the Endowed Chair of Applied Artificial Intelligence, Oldenburg University.

REFERENCES

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *International Conference on Learning Representations*, 2023.

Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pp. 339–349. Springer, 2021.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

Shunjie Dong, Zixuan Pan, Yu Fu, Dongwei Xu, Kuangyu Shi, Qianqian Yang, Yiyu Shi, and Cheng Zhuo. Partial unbalanced feature transport for cross-modality cardiac image segmentation. *IEEE Transactions on Medical Imaging*, 42(6):1758–1773, 2023.

Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16549–16559, 2021.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Minghao Han, Linhao Qu, Dingkang Yang, Xukun Zhang, Xiaoying Wang, and Lihua Zhang. Mscpt: Few-shot whole slide image classification with multi-scale and context-focused prompt tuning. *arXiv preprint arXiv:2408.11505*, 2024.

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*, 29(9):2307–2316, 2023.

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.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Kwanyoung Kim, Yujin Oh, and Jong Chul Ye. Zegot: Zero-shot segmentation through optimal transport of text prompts. *arXiv preprint arXiv:2301.12171*, 2023.

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*, pp. 14318–14328, 2021.

Honglin Li, Chenglu Zhu, Yunlong Zhang, Yuxuan Sun, Zhongyi Shui, Wenwei Kuang, Sunyi Zheng, and Lin Yang. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7454–7463, 2023.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers*, 2021.

Zhuowei Li, Long Zhao, Zizhao Zhang, Han Zhang, Di Liu, Ting Liu, and Dimitris N Metaxas. Steering prototypes with prompt-tuning for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2523–2533, 2024.

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*, pp. 19830–19839, 2023.

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.

Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19764–19775, 2023.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5206–5215, 2022.

Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical image analysis*, 33:170–175, 2016.

Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Hibou: A family of foundational vision transformers for pathology. *arXiv preprint arXiv:2406.05074*, 2024.

Duy MH Nguyen, Nghiem T Diep, Trung Q Nguyen, Hoang-Bao Le, Tai Nguyen, Tien Nguyen, TrungTin Nguyen, Nhat Ho, Pengtao Xie, Roger Wattenhofer, et al. Logra-med: Long context multi-graph alignment for medical vision-language model. *arXiv preprint arXiv:2410.02615*, 2024a.

Duy MH Nguyen, An T Le, Trung Q Nguyen, Nghiem T Diep, Tai Nguyen, Duy Duong-Tran, Jan Peters, Li Shen, Mathias Niepert, and Daniel Sonntag. Dude: Dual distribution-aware context prompt learning for large vision-language model. *Asian Conference on Machine Learning (ACML))*, 2024b.

Huy Nguyen, Khang Le, Quang Nguyen, Tung Pham, Hung Bui, and Nhat Ho. On robust optimal transport: Computational complexity and barycenter computation. In *Advances in NeurIPS*, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of Sinkhorn algorithm. In *International Conference on Machine Learning*, pp. 7673–7682. PMLR, 2020.

Linhao Qu, Kexue Fu, Manning Wang, Zhijian Song, et al. The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. *Advances in Neural Information Processing Systems*, 36, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Jeongun Ryu, Aaron Valero Puche, JaeWoong Shin, Seonwook Park, Biagio Brattoli, Jinhee Lee, Wonkyung Jung, Soo Ick Cho, Kyunghyun Paeng, Chan-Young Ock, et al. Ocelot: overlapped cell on tissue dataset for histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23902–23912, 2023.

Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. *Handbook of Numerical Analysis*, 24:407–471, 2023.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer-based correlated multiple instances learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Jiangbo Shi, Lufei Tang, Yang Li, Xianli Zhang, Zeyu Gao, Yefeng Zheng, Chunbao Wang, Tieliang Gong, and Chen Li. A structure-aware hierarchical graph-based multiple instance learning framework for pt staging in histopathological image. *IEEE Transactions on Medical Imaging*, 42(10): 3000–3011, 2023.

Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. 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.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4078–4087, 2023.

Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. 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.

The Cancer Genome Atlas (TCGA). Genomic Data Commons Data Portal (GDC). `https://portal.gdc.cancer.gov/projects/TCGA-BRCA`. Accessed 07 Jul. 2023.

Masayuki Tsuneki and Fahdi Kanavati. Inference of captions from histopathological patches. In *International Conference on Medical Imaging with Deep Learning*, pp. 1235–1250. PMLR, 2022.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on computer vision*, pp. 10682–10691, 2019.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, pp. 1–8, 2024.

Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23438–23448, 2024.

Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15028–15038, 2021.

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*, pp. 18802–18812, 2022a.

Yue Zhang, Hongliang Fei, Dingcheng Li, Tan Yu, and Ping Li. Prompting through prototype: A prototype-based prompt learning on pretrained vision-language models. *arXiv preprint arXiv:2210.10841*, 2022b.

Cairong Zhao, Yubin Wang, Xinyang Jiang, Yifei Shen, Kaitao Song, Dongsheng Li, and Duoqian Miao. Learning domain invariant prompt for vision-language models. *IEEE Transactions on Image Processing*, 2024.

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.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.