Supplementary Material
Multi-Resolution Pathology-Language Pre-training Model with Text-Guided Visual Representation

Shahad Albastaki[1], Anabia Sohail[1], Iyyakutti Iyappan Ganapathi[1], Basit Alawode[1], Asim Khan[1], Sajid Javed[1,*], Naoufel Werghi[1], Mohammed Bennamoun[3], Arif Mahmood[2]
[1]Department of Computer Science, *ARIC, Khalifa University of Science and Technology, UAE
[2]Information Technology University of the Punjab, Pakistan, [3]University of the Western Australia

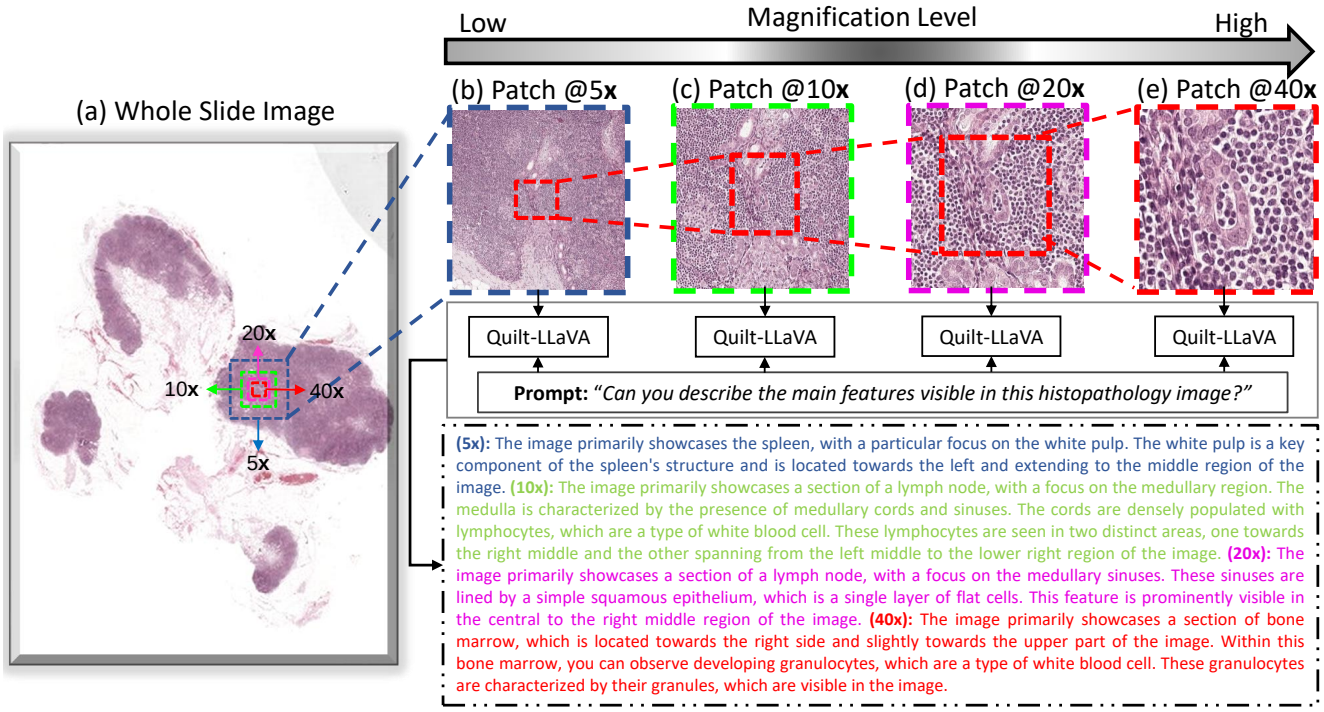
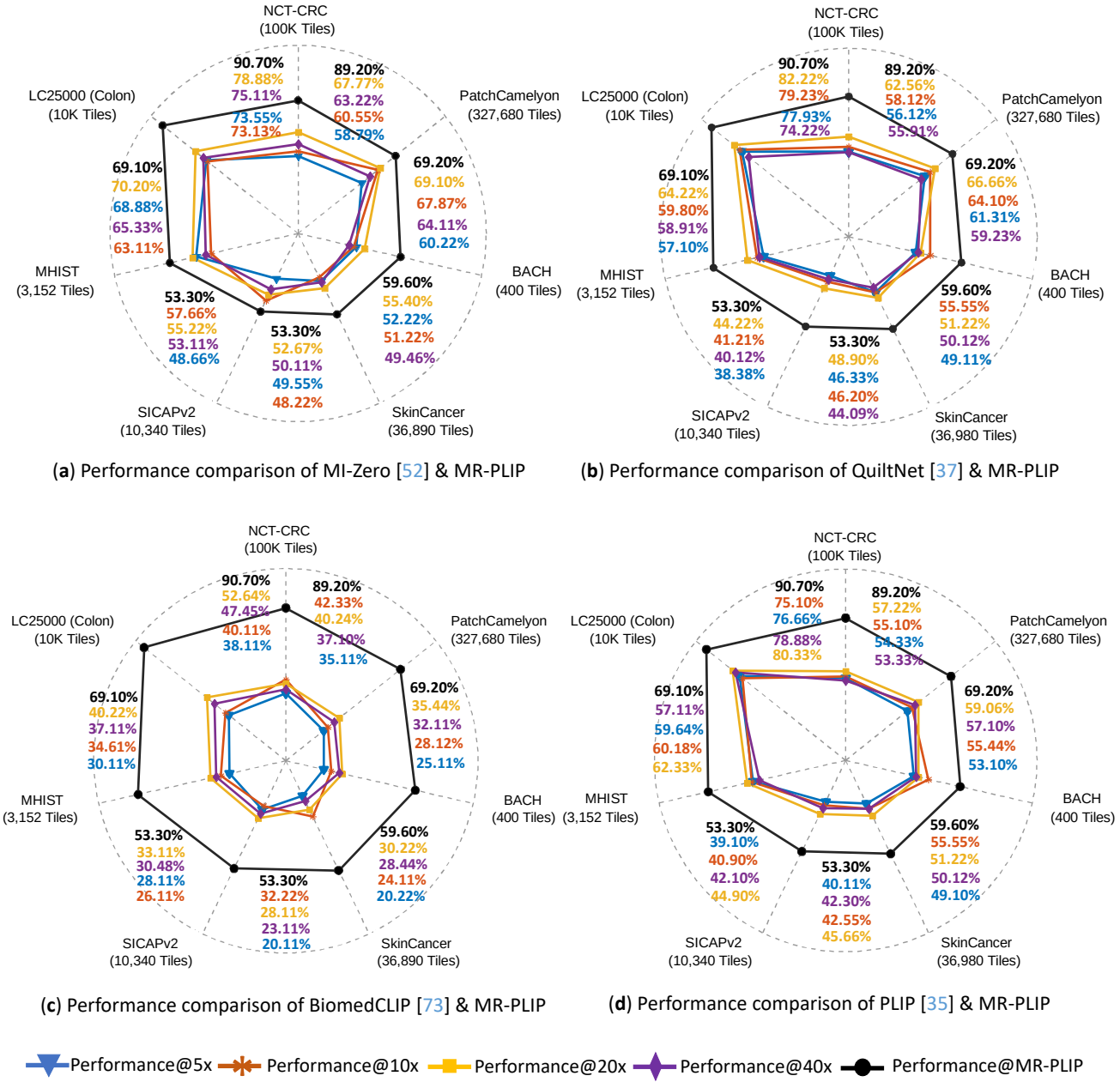


Figure 1. Example of multi-resolution analysis of a histology image extracted from input WSI **(a)** using the Quilt-LLaVA [59]. Exemplar histology patches **(b)-(e)** are shown at different magnifications, demonstrating how higher magnification (5× to 40×) shifts focus from contextual to detailed information. Textual descriptions generated by Quilt-LLaVA vary, reflecting the change in textual details observed at each magnification level from 5× to 40×.

Table of Contents



(**a**) Performance comparison of MI-Zero [52] & MR-PLIP

(**b**) Performance comparison of QuiltNet [37] & MR-PLIP

(**c**) Performance comparison of BiomedCLIP [73] & MR-PLIP

(**d**) Performance comparison of PLIP [35] & MR-PLIP

Figure 2. Zero-shot tile-based classification performance in terms of accuracy of SOTA VL models including (a) MI-Zero [52], (b) QuiltNet [37], (c) BiomedCLIP [73], and (d) PLIP [35] on testing splits of seven independent datasets. Both models are pre-trained on the TCGA dataset using patches from 20k WSIs. In each experiment, the pre-trained vision-language encoders are fine-tuned on a fixed magnification level 5×, 10×, 20×, or 40×. Performance variations across different magnification levels show the need for a multi-resolution VL model in computational pathology for improved generalization capabilities.

1. MR-PLIP Model: More Insights

In clinical diagnostics, expert pathologists often analyze the WSI to predict the outcomes of diseases by inspecting it from multiple resolution levels. The multi-resolution analysis of the WSIs assists expert pathologists to better analyze the tumor micro-environment by looking at the surrounding tissue/cellular structures [1, 11, 18, 25, 34, 74]. For example, pathologists look at the global architectural composition of the tissue sample and analyze the context of each tis-

sue component, including cancer, to identify the presence of both healthy and cancerous tissues. Additionally, they zoom in into each region of interest, where the tissue is examined at a high resolution, to obtain the details of the cancer cells, and characterize the tumor based on its local cellular composition. Another example where pathologists take advantage of both context and details is the spatial distribution of immune cells, which may be detected in the presence of inflammation inside the tumor or the stromal compartment of the cancer regions, as well as in specific clustered groups called tertiary lymphoid structures, which may develop in response to cancer as shown in Figs. 1.

The above multi-scale analysis is crucial, as it involves the integration of both overarching (i.e., viewing the WSI at the lowest level of magnification) and detailed (i.e., viewing the WSI at the highest level of magnification) viewpoints [14, 21, 22, 33]. Such a thorough approach enables pathologists to accurately distinguish between various types of cancer, such as differentiating invasive ductal carcinoma from invasive lobular carcinoma, as well as identifying tumor-infiltrating lymphocytes [6, 14].

Most contemporary VLMs in histopathology primarily use histology images extracted from WSI of a single resolution, which might restrict their capability to adequately convey the essential broad and detailed perspectives for optimal analysis [35, 51, 52]. An illustration of this, as shown in Figs. 1, is provided by the SOTA VLM, Quilt-LLaVA [59], which demonstrates how, with increasing magnification levels, the amount of textual descriptions derived from the input histology patch decreases. This decline is attributed to the loss of contextual information at higher magnifications. Additionally, pivotal cues, such as those indicating invasive lymph node, may only be visible at specific magnifications, highlighting the Quilt-LLaVA model's considerable dependency on certain magnification levels for generating accurate textual descriptions, a limitation that might be seen as a drawback.

To explore the multi-resolution generalization abilities of the SOTA methods, we fine-tuned SOTA CPath models, including PLIP [35], BiomedCLIP [73], MI-Zero [52], and QuiltNet [37], across magnification levels of 5×, 10×, 20×, and 40×, using 20,000 WSIs (comprising 34 million patches) from the TCGA dataset [71]. These models are assessed through zero-shot settings across seven benchmark datasets for tile-based classification, as depicted in Fig. 2. Excluding our top-scoring mode MR-PLIP, Fig. 2 shows that 20× is virtually the best, coming first in 13 out of 14 trials. The 10× consistently ranks either first or second in 8 out of 14 trials, while the extreme magnifications of 5× and 40× typically land in the last or third position in 10 out of 14 trials. These statistics highlight that the magnifications 20× and 10×, striking a balance between detail and context, achieve optimal performance. Our intuition is that integrating the 5× and 40× alongside the 20× and 10× in VL models will further leverage the complementary of the four magnification levels. We suggest, therefore, that synchronizing visual-textual concepts across multiple resolutions enhances their efficacy for diverse CPath applications and their overall generalization.

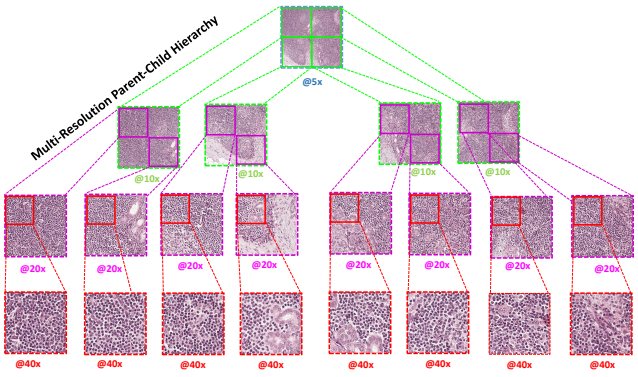


Figure 3. Parent-child relationships in the visual bag ($B^v_{i,j}$). Our loss ($\mathcal{L}_{MRTVA}$) is minimized while preserving the parent-child relationships.

2. Parent-Child Hierarchy

To clarify the approach used to preserve the hierarchical structure in our curated histopathology dataset, particularly when aligning text-guided visual features across different resolution levels as outlined in Eq. (3) of the main manuscript, we focus on maintaining the integrity of the parent-child relationship, as depicted in Fig 3. Specifically, the alignment of text-guided visual features is strictly between parents and their direct offspring. Within the visual bag $B^v_{i,j}$, patches lacking a parent-child linkage do not share visual content, rendering their alignment irrelevant. In our MR-PLIP model, alignment is confined to parents with their immediate children and vice versa, enabling the model to assimilate contextual and intricate details across different resolution levels. Alignments between grandparents or grandchildren are omitted to avoid confusion from minimally overlapping content.

3. Multi-modal Encoder for Text-guided Visual Representation

Our multi-modal encoder's architecture incorporates elements from the frameworks presented in [45, 48, 72], which originally were not applied to Computational Pathology (CPath). Here, we explicitly adapt and merge their methodologies for the first time to suit the specific needs of the CPath domain, as demonstrated in Fig. 4, showing the encoder's architecture. This model incorporates the identical unimodal encoders for text (using the QuiltNet model)

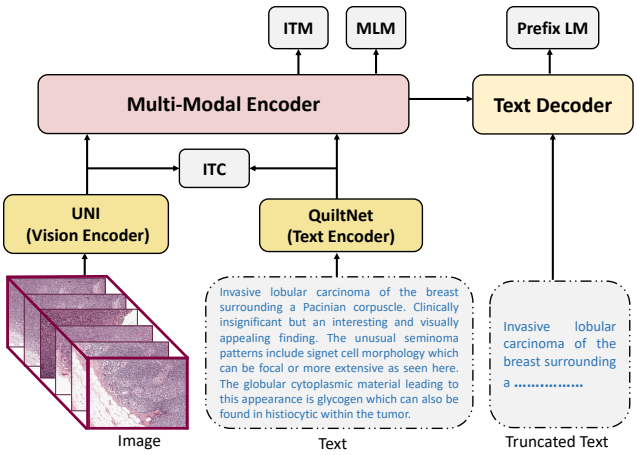


Figure 4. The architecture of our multi-modal encoder for estimating text-guided visual feature representation (see Sec. 3.4 in the main manuscript).

and images (using the UNI model), as outlined in the main manuscript, along with a multi-modal encoder for integrating text-guided visual features and a text decoder for generating text. In particular, the vision encoder processes a multi-resolution histology patch, $p^r_{i,j}$, converting it into a visual representation, $v^r(i,j)$. Similarly, the text encoder converts the corresponding textual descriptions, $t^r_{i,j}$, of $p^r_{i,j}$ into textual features, $w^r_{i,j}$. To effectively merge multi-modal data while retaining the integrity of single-mode information, we initially combine the image and text features derived from the unimodal encoders as per references [45, 48]. To achieve this, the Image-Text Contrastive (ITC) loss is employed to synchronize the unimodal outputs from both the visual and text encoders. Subsequently, a cross-modal network with skip connections is utilized to merge the visual and textual data efficiently, applying the Image-Text Matching (ITM) and Masked Language Modeling (MLM) losses [24, 48] for effective fusion. The decoder, informed by the integrated image and prefix sub-sequence representation, is trained using the Prefix Language Modeling (PLM) loss to complete the caption generation [13].

4. Pre-Training Baseline Objectives

During the pre-training process, we perform four pre-training tasks: Image-Text Contrastive Learning ($\mathcal{L}_{ITC}$), Image-Text Matching ($\mathcal{L}_{ITM}$), Masked Language Modeling ($\mathcal{L}_{MLM}$), and Prefix Language Modeling ($\mathcal{L}_{PLM}$). First, the ITC task is employed to align the unimodal representations of images and texts. Then, the ITM and MLM tasks are used for learning the multi-modal representation. Based on the image-language representations produced by the multi-modal encoder, the decoder is then trained with PLM loss to perform text-completion tasks.

4.1. Image-Text Contrast (ITC):

In line with [46, 48], this task is used to align the unimodal encoders. Specifically, we calculate the softmax-normalized similarities between image-to-text and text-to-image, incorporating memory queues as described in MoCo [20] to expand the pool of negative samples during the learning process. For each patch $p^r_{i,j}$, we use its visual feature vector $v^r_{i,j}$ along with its corresponding k_o positive words to generate the textual feature representation $w^r_{i,j}$. Through the application of two projector networks [48], these visual and textual features are then transformed into $v^{r'}_{i,j}$ and $w^{r'}_{i,j}$ respectively.

Formally, the Image-Text Contrastive (ITC) loss is then calculated as:

$$\mathcal{L}_{i2t} = -\frac{1}{M}\sum_{m=1}^{M}\log\frac{\exp(s(v^{r'}_{i,j}, w^{r'}_{i,j})/\tau)}{\sum_{m=1}^{M}\exp(s(v^{r'}_{i,j}, w^{r'}_{i,m})/\tau)}, \quad (1)$$

$$\mathcal{L}_{t2i} = -\frac{1}{M}\sum_{m=1}^{M}\log\frac{\exp(s(w^{r'}_{i,j}, v^{r'}_{i,j})/\tau)}{\sum_{m=1}^{M}\exp(s(w^{r'}_{i,j}, v^{r'}_{i,m})/\tau)}, \quad (2)$$

$$\mathcal{L}_{ITC} = \frac{1}{2}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}), \quad (3)$$

where $v^{r'}_{i,m}$ contain both the positive and the negative visual samples for text representation $w^r_{i,j}$ and $w^{r'}_{i,m}$ are the positive and negative text samples for visual feature $v^r_{i,j}$.

4.2. Image-Text Matching (ITM):

This task aims to predict whether an image and a text are paired or not based on the multi-modal representation [45, 48]. In this loss, we maximize the log-likelihood of predicting a positive and negative pair, given the visual and textual descriptions $v^{r'}_{i,j}$ and $w^{r'}_{i,j}$.

$$\mathcal{L}_{itm} = -\mathbb{E}_{(v^{r'}_{i,j}, w^{r'}_{i,j})}\log p(y|v^{r'}_{i,j}, w^{r'}_{i,j}) \quad (4)$$

where $y \in \{+,-\}$ is the predicted label during the contrastive learning.

4.3. Masked Language Modeling (MLM):

In this task, tokens $w^{r'}_{i,j,b}$ are masked, and the model is tasked with predicting these masked tokens by leveraging the multi-modal representation [24]. The loss for this masked language modeling is defined as per references [24, 45, 48]:

$$\mathcal{L}_{mlm} = -\mathbb{E}_{(v^{r'}i,j,w^{r'}_{i,j,b})}\log p(w^{r'}_{i,j,b}|v^{r'}_{i,j}, w^{r'}_{i,j,/b}) \quad (5)$$

where $w^{r'}_{i,j,/b}$ are the non-masked key words.

4.4. Prefix Language Modeling (PLM):

This pretext task requires the model to complete the truncated texts based on the given image and prefix sequence of truncated texts [45, 48, 49]. The model can be trained by maximizing the likelihood of the truncated text in an auto-regressive manner. Formally, the prefix language modeling loss is calculated as [45, 49]:

$$\mathcal{L}_{plm} = -\mathbb{E}_{(v^{r'}_{i,j} w_b)} \left(\sum_{l=l_p}^{L} \log p\Big(w^{r'}_{i,j,l}) | w^{r'}_{i,j,[l_p,l]}, w^{r'}_{i,j,<l_p}, v^{r'}_{(i,j)}\Big)\right), \tag{6}$$

where L denotes the total number of words and l_p is the length of a pre-fix sequence of keywords that is randomly selected.

5. Additional Training and Implementation Details

The TCGA dataset is renowned for being one of the largest publicly available histopathology collections, covering a broad spectrum of cancer morphologies and subtypes across key organs [71]. We included WSIs from 21 major primary cancer sites in TCGA, covering the adrenal gland, bile duct, bladder, bone marrow, breast, testis, pleura, cervix, eye, head and neck, stomach, uterus, thyroid, pancreas, esophagus, ovary, liver, endometrium, thymus, skin, and larynx. Testing was conducted on novel organs and cancer subtypes—including colorectal, kidney, prostate, and brain—as well as seen types such as breast, lung, bone, skin, and NSCLC, without any overlap with the training data. Results on both seen and unseen cases demonstrated superior performance compared to SOTA methods.

For the customization of the MR-PLIP algorithm to this dataset, we refined it through the creation of 34 million multi-resolution histology patches, each measuring 512 × 512 pixels. Initially, we applied the Otsu method [54] for WSI thresholding, then proceeded to extract patches at designated magnification levels (as detailed in Sections 3.1 and 3.2 in the main manuscript). These patches were chosen based on a criterion ensuring at least 70% tissue coverage to highlight pertinent histological details.

Our pre-training methodology integrated various architectures, including both domain-specific and general models. The fine-tuning process for our MR-PLIP model involved initializing it with diverse sets of weights for image and text encoders. An ablation study was conducted to evaluate the MR-PLIP model's performance across different setups (refer to Table 4). We initialized the domain-specific vision encoder with UNI (ViT-L/16), pre-trained on histopathology data [19], and the domain-specific text encoder with the first six layers of the QuiltNet model [37], a GPT-2 adaptation with a context length of 77 (GPT-2/77) [37]. The multi-modal encoder was similarly initialized using the latter six layers of the pre-trained QuiltNet model (GPT-2/77). The MR-PLIP model was pre-trained over 50 epochs, with batch sizes set to 32 across six NVIDIA A100 GPUs. The AdamW optimizer [3] was used for optimization, featuring a weight decay of 0.02 and beta values of (0.9, 0.98). The learning rate experienced an initial ramp-up to 5e-5 across the first 1000 iterations, subsequently following a cosine decay pattern. Through empirical testing during pre-training, we established the optimal number of positive keywords (k_o) at 9, and set the Masked Language Modeling (MLM) mask ratio to 15% [48].

During the fine-tuning phase of our MR-PLIP model, we used various combinations of image and text encoders to enhance its performance, detailed as follows:

1. In alignment with SOTA methods [35, 37, 52, 73], we fine-tuned the baseline CLIP model [57] using a ViT-B/16-224 [26] as the image encoder and GPT-2/77 [56] as the text encoder.
2. Recognizing the baseline CLIP's training on out-of-domain paired data, we also fine-tuned the MR-PLIP model with a pathology domain-specific pre-trained PLIP [35] , using PLIP-ViT-B/32-224 as the image encoder and GPT-2/347 as the text encoder.
3. Following the approach of MI-zero and Biomed-CLIP [73], we fine-tuned the MR-PLIP model using BioClinicalBert/512 [5] and PubMedBERT/256 [31] as text encoders, alongside CTransPath/224 [68] as the image encoder. BioClinicalBert and PubMedBERT, both non-pathology text encoders, are trained on biomedical and clinical corpora, such as PubMed abstracts and MIMIC [40]. CTransPath is trained through self-supervised learning on 15.5 million unlabeled histology patches, with both encoders using ViT-B/16.
4. MR-PLIP was also fine-tuned using BioClinicalBert/512 as the text encoder and PLIP-ViT-B/32-224 as the in-domain image encoder.
5. Furthermore, we fine-tuned MR-PLIP using CTransPath/224 as the in-domain image encoder and PLIP-GPT/347 as the in-domain text encoder.
6. Additionally, we fine-tuned MR-PLIP using ViT-B/16 as the in-domain image encoder from the QuiltNet and GPT-2/77 as the in-domain text encoder of the QuiltNet model [37].

These experiments were conducted using inference time prompts similar to [51] for a fair comparison. Throughout this paper, our reported experiments predominantly used the UNI (ViT-L/16) [19] as the image encoder and GPT-2/77 as the text encoder from the QuiltNet model [37]. Please see Table 4 for comparison.

6. Evaluation Metrics

For evaluating performance across different CPath tasks [35, 51], we use a variety of metrics. These include the weighted average F_1 score, balanced accuracy, dice score, precision, recall, multi-class Panoptic Quality (mPQ), Recall@1 (R@1), Recall@50 (R@50), and Recall@200 (R@200). Consistent with prior VLMs studies in CPath, the weighted average F_1 score and balanced accuracy are applied to gauge performance in tile-level and WSI-level classification tasks. For segmentation tasks, we measure using dice score, precision, and recall. The mPQ metric is specifically used for nuclei instance segmentation tasks, while R@1, R@50, and R@200 metrics are dedicated to evaluating the efficacy of cross-modal retrieval tasks.

7. Downstream Histopathology Datasets

We performed five different computational pathology (CPath) tasks, including ***tile-level*** classification, ***WSI-level*** classification, ***cross-modal retrieval***, ***WSI segmentation***, and ***nuclei segmentation***. To evaluate our MR-PLIP model on these tasks, we used 26 independent datasets. For fair comparisons with SOTA methods [35, 37, 41, 52], we employed the official testing splits of the datasets for zero-shot evaluation. For linear probing and fine-tuning experiments, we employed the official training and testing splits of each dataset. The details of each dataset are outlined below:

7.1. Tile-level Classification Datasets

We used 15 independent datasets, consisting of a wide range of tissue images from various resolution levels across different cancer types and tissues. These datasets include Databiox [14], which focuses on invasive ductal carcinoma from 124 patients, BACH [36] for breast cancer analysis from 500 WSIs, PatchCamelyon [66] for identifying normal and metastatic tumor tissues from 400 WSIs, WILDS-CAM17 [9, 43] for classifying breast metastasis, UniToPatho [10] for colorectal polyp classification, and Osteo [7] for osteosarcoma from 40 WSIs. Additionally, SkinCancer dataset [44] provides 36,890 skin tissue patches for identifying various skin conditions, MHIST[70] for colorectal polyps analysis, RenalCell [16] for studying clear-cell renal cell carcinoma, and several others focusing on specific cancers like lung and colon cancer, as well as datasets like DigestPath [23] for colonoscopy analysis, SICAP [62] for prostate cancer Gleason pattern classification, and WSSS4LUAD [32] for lung adenocarcinoma. These datasets, with their specific focuses, resolutions, and classifications, provide a comprehensive resource for validating and fine-tuning the MR-PLIP model's capability in accurately classifying a wide array of histopathological images, demonstrating its adaptability and precision across different domains within pathology.

- **Databiox (3 Classes):** is an invasive ductal carcinoma dataset collected from pathological biopsy samples of 124 patients. This dataset comprises 922 samples, corresponding to 2100×1574 and 1276×956 pixels. Each sample is captured at four different levels of magnification, including $4\times$, $10\times$, $20\times$, and $40\times$. The samples are annotated into Grade I (well-differentiated), Grade II (moderately differentiated), and Grade III (poorly differentiated).
- **BACH (4 Classes):** is a breast cancer dataset containing 500 large tiles, each with 2048×1536 pixels captured at $40\times$ magnification and sampled from 500 WSIs. The dataset is classified into four different tissue types, including normal, benign, in-situ carcinoma, and invasive carcinoma. The official training (320 tiles) and testing (80 tiles) splits are provided.
- **PatchCamelyon (2 Classes):** is a breast cancer dataset containing normal and metastatic tumor tissues. This dataset is collected from 400 WSIs, containing 327,680 H&E stained histology images with 96×96 pixel tiles. The samples are extracted from lymph node sections at $10\times$ magnification level to provide an increased field of view. The official training, validation, and testing splits contain 262,144, 32,768, and 32,768 histology images.
- **WILDS-CAM17 (2 classes):** is a patch-based breast metastasis detection dataset based on CAM17 dataset, with folds created by WILDS for testing the models' robustness under distribution shift. The dataset consists of 417,894 patches, each with 96×96 pixels extracted from WSIs of breast cancer metastases in lymph node sections. The patch label refers to whether the patch contains a tumor or is normal. For training and evaluation, we used the official train–validation–test folds provided by WILDS. The training set contains 302,436 patches from three hospitals, and the model is evaluated on 34,904 validation patches and 80,554 testing patches. We resized images at 224×224 pixels and evaluated SOTA methods.
- **UniToPatho (6 classes):** is a colorectal polyp classification dataset containing 9,536 each with $1,812 \times 1,812$ pixels at 0.44 μm/pixel annotated and extracted from 292 WSIs. The dataset contains six tissue types, including normal (950 patches), hyperplastic polyp (545 patches), tubular adenoma with high-grade dysplasia (454 patches), tubular adenoma with low-grade dysplasia (3,618 patches), tubulovillous adenoma with high-grade dysplasia (916 patches), and tubulovillous adenoma with low-grade dysplasia (2,186 patches). The official train set consists of 6,270 patches, while the testing set contains 2,399 patches. We resized images at 224×224 pixels and evaluated SOTA methods.
- **Osteo (3 Classes):** dataset focuses on osteosarcoma and contains 1,144 patches of size 1024×1024 pixels. The dataset is collected from 40 heterogeneous WSIs at $10\times$

magnification level. The dataset contains three distinct classes, including tumor, non-tumor, and necrotic tumor. The official train and test splits are provided with a ratio of 80:20.

- **SkinCancer (16 Classes):** comprises 36,890 skin tissue patches extracted from 386 patients, each with 395×395 pixels. The patches are captured at $10\times$ magnification from patients with basal cell carcinoma, squamous cell carcinoma, naevi, and melanoma. The tiles are categorized into 16 distinct categories, including chondral tissue, dermis, elastosis, epidermis, hair follicle, skeletal muscle, necrosis, nerves, sebaceous glands, subcutis, eccrine glands, vessels, BCC, SqCC, naevi, melanoma. The official train (88971), validation (72348), and testing (28039) splits are provided.
- **MHIST (2 Classes):** is a colorectal polyps dataset that contains 3,152 tissue patches of size 224×224 pixels extracted at $40\times$ magnification level from 328 WSIs. The tiles are annotated into two classes, including hyperplastic polyps and sessile serrated adenomas. For training, 2,175 tiles are utilized, while 977 tiles are reserved for testing.
- **RenalCell (5 Classes):** dataset contains histology patterns of clear-cell renal cell carcinoma. The dataset consists of 52,713 H&E-stained images with 300×300 pixels captured at $40\times$. The dataset is annotated into five distinct classes, including red blood cells, renal cancer, normal tissue, torn adipose necrotic tissue, and muscle fibrous stroma blood vessels.
- **NCT-CRC (9 Classes):** is a colorectal cancer dataset comprising H&E stained images encompassing nine distinct classes including Adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium. Tissue patches, corresponding to 224×224 pixels, are extracted at $20\times$ magnification level. The training dataset comprises 100K patches extracted from 86 patients, while the testing set consists of 7,180 images extracted from 50 patients.
- **LC25000Lung (3 Classes):** is a lung cancer dataset comprising 25K H&E stained images corresponding to 768×768 pixels extracted from 750 patients. The dataset encompasses three classes, including lung adenocarcinoma, benign lung, and lung squamous cell carcinoma. The official training and testing splits are provided with a ratio of 80:20.
- **LC25000Colon (2 Classes):** is a colon cancer dataset containing H&E stained images each of size 768×768 pixels extracted from 500 patients. This dataset contains two classes: benign colon tissue and colon adenocarcinomas. The official training and testing splits are provided with a ratio of 80:20.
- **DigestPath (2 Classes):** is a colonoscopy dataset containing H&E 660 tissue images. Similar to PLIP, we performed tile-based zero-shot classification for Tumor Vs. Normal on the testing split containing 18814 images.
- **SICAP (4 Classes):** is a prostate cancer dataset tailored for Gleason pattern classification. It encompasses 512×512 pixels tiles extracted from 155 WSIs. The official training split comprises 9,959 images sourced from 124 WSIs, while the testing split includes 2,122 images from 31 WSIs. The dataset encompasses four labels, indicating the primary Gleason pattern (3, 4, or 5) or noncancerous (NC).
- **WSSS4LUAD (2 Classes):** is a lung adenocarcinoma dataset containing tiles of 200×500 pixels. It encompasses three distinct classes: tumor, tumor-associated stroma, and normal. We conducted a binary classification of tumor vs. normal. The training dataset comprises 7,063 images, while the testing set comprises 3,028 images (2,015 tumors, 1,013 normal).

7.2. Tumor Subtype Classification Datasets

We used eight distinct datasets for tumour subtyping at the WSI level: CAMELYON16 (CAM16) [12], CAMELYON-17 (CAM17) [9], RCC-DHMC [75], BRCA-BRACS [15], HunCRC [55], PANDA [17], EBRAINS [58, 65] and NSCLC-CPTAC [27], all of which provide only slide-level labels. Following the approach of CONCH [51] and MI-Zero [52], we employed a top-K pooling operator for WSI classification task. CAM16 focuses on breast cancer, specifically the detection of lymph node metastasis from 400 gigapixel WSIs, with a training split of 270 WSIs and a testing split of 130 WSIs, where 159 are normal, and 111 contain tumor regions. CAM17, another breast cancer dataset, originates from sentinel lymph node sections of 200 patients, comprising 1000 WSIs with 5 slides per patient, labeled for metastasis vs. normal, with the model evaluated on 500 WSIs for both training and testing. CPTAC dataset, consisting of 1091 WSIs of different patients, focuses on LUng ADenocarcinoma (LUAD) and LUng Squamous cell Carcinoma (LUSC) and offers images at different resolutions, providing a comprehensive set of data for validating WSI-level classification performance. EBRAINS is a dataset for 30-way fine-grained brain tumor subtyping. RCC-DHMC is a Renal Cell Carcinoma (RCC) subtyping dataset. HunCRC is a colorectal cancer screening dataset. BRCA-BRACS is a breast cancer subtyping dataset while PANDA is a prostate cancer dataset for gleason scoring.

- **CAMELYON16 (CAM16) (2 Classes)** is a breast cancer dataset designed for detecting lymph node metastasis using gigapixel WSIs. It comprises a total of 400 WSIs with only slide-level labels provided. The official training split consists of 270 WSIs, while the testing split contains 130 WSIs. Within the training set, there are 159 normal WSIs, and 111 WSIs contain tumor regions indicating breast cancer metastasis.

- **CAMELYON17 (CAM17) (2 Classes)** is a breast cancer dataset generated from a sentinel lymph node section of the Breast from 200 patients. The Camelyon17 dataset contains 1000 WSIs with 5 slides per patient. The WSI is labeled as metastasis vs. normal. The proposed model is evaluated on the official train (500 WSIs) and test set (500 WSIs) splits.
- **NSCLC-CPTAC (2 Classes)** is a Non-Small Cell Lung Carcinoma (NSCLC) subtyping based on CPTAC contains two classes including LUAD and LUSC. We excluded slides that were frozen tissue, nontumor tissue, or were not labeled as having acceptable tumor segments, which resulted in 1,091 slides (578 LUAD and 513 LUSC). For training and evaluation, we divided datasets into train-validation-test folds with 80:10:10 ratio and 872:109:109 slides. For zero-shot classification, we used 109 testing WSIs.
- **EBRAINS [58, 65] (30 classes)** is a dataset consists of H&E histopathology WSIs of brain tissue selected from The Digital Brain Tumour Atlas an open histopathology resource. Similar to the CONCH [52], we used a subset of 2,319 WSIs out of 3,114 WSIs and defined a 30-way fine-grained brain tumor subtyping task limited to diagnostic labels that have at least 30 slides. For the supervised dataset, we performed a 50–25–25 split for training (1,151 slides), validation (595 slides) and testing (573 slides). For the zero-shot test set, we used the testing split of 573 slides. The WSI counts for each class in the dataset were also set according to the CONCH [52].
- **RCC-DHMC (5 classes)** is a Renal Cell Carcinoma (RCC) subtyping dataset consisting of 563 RCC H&E diagnostic histopathology WSIs. The dataset contains six cancer subtypes, including primary Clear Cell Renal Cell Carcinoma (CCRCC) (344 slides), Papillary Renal Cell Carcinoma (PRCC) (101 slides), CHromophobe RCC (CHRCC) (23 slides), Renal Oncocytomas (ROCY)(66 slides), and benign cases (29 slides). Similar to UNI [19], for training and evaluation, we used a modified configuration of the train–validation–test folds with a 70:4:26 ratio (393:23:147 slides), with eight CHRCC cases moved from the test to the training fold due to CHRCC being absent in the training fold.
- **HunCRC (4 Classes)** is colorectal cancer screening dataset containing of 200 H&E diagnostic histopathology WSIs of colorectal biopsies. Similar to UNI [19], we employed a 4-way coarse-grained subtyping task using the categories of normal (10 slides), non-neoplastic lesion (38 slides), CRC (46 slides), and adenoma (106 slides), in which the ground-truth label was set by the study's pathologist. For training and evaluation, we split the dataset into 50:25:25 train–validation–test folds (158:21:21 slides).
- **BRCA-BRACS (3 classes)** is a BReast CAncer (BRCA) dataset containing 547 breast carcinoma H&E WSIs from 187 patients sourced from the breast carcinoma subtyping task. The dataset contains 3 classes, including benign, atypical, and malignant tumor labels. For training and evaluation, we used the official train–validation–test folds with a 72:12:16 ratio (395:65:87 WSIs).
- **PANDA (6 classes)** is the International Society of Urological Pathology (ISUP) grading task derived from the PANDA challenge. It consists of 10,616 prostate cancer core needle biopsies of the prostate. Each slide is assigned an ISUP score that defines prostate cancer grade (6-class grading task). We removed the noisy labels and considered 9,555 slides only in which Grade 0 (G0) contains 2,603 WSIs, G1 contains 2,399 WSIs, G2 contains 1,209WSIs, G3 contains 1,118 WSIs, G4 contains 1,124 WSIs, and G5 contains 1,102). For training and evaluation, we employed 80:10:10 train–validation–test folds (7,647:954:954 WSIs).

7.3. Histology Image Segmentation Datasets

We used three distinct datasets for histolog image segmentation tasks, including DigestPath, SICAP, and TIGER [60].

- **DigestPath:** focuses on colonoscopy H&E tissue sections, comprising 660 images with pixel-level lesion annotations for colorectal cancer provided for 250 testing images from 93 patients.
- **SICAP:** This dataset aimed at prostate cancer, facilitates Gleason pattern classification with 31 WSIs in the testing split for tumor vs. normal tissue segmentation, and 124 WSIs in the training split.
- **TIGER:** dataset, dedicated to tumor-infiltrating lymphocytes (TILs) score prediction in H&E stained breast cancer images, includes 195 WSIs from as many patients. These WSIs are annotated for semantic segmentation to identify regions such as invasive tumor, tumor-associated stroma, in-situ tumor, and others. For our study, we simplified the classification into three classes: tumor, stroma (merging tumor-associated and inflamed stroma), and background, by combining invasive and in-situ tumors to denote the tumor region, with all other regions marked as background.

7.4. Nuclei Segmentation Datasets

For the nuclei segmentation task, we follow the same settings as proposed in DINOSSLPath [41] by employing a HoverNet [30] baseline model. We used two publicly available datasets: PanNuKe [29] and CONSEP [30].

- **CONSEP:** This dataset focuses on diverse nuclei segmentation and classification across six nuclear types, containing 41 image tiles of 1000×1000 pixels each, captured at $40\times$ magnification, with 26 images designated for training and 14 for testing.

Table 1. Ablation study of MR-PLIP's zero-shot classification performance (weighted F_1 score) using different multi-resolution image-text pairs. The best resolution combinations for each dataset are highlighted.

Resolutions	CAM16	CPTAC	SICAP	DigestPath	Databiox	NCT-CRC
5×,10×	0.566	0.706	0.421	0.792	0.421	0.732
20×,40×	0.631	0.805	0.529	0.887	0.472	0.835
5×,10×, 20×	0.599	0.771	0.469	0.883	0.518	0.807
10×,20×,40×	0.621	0.843	0.502	0.891	0.506	0.834
5×,10×,20×,40×	**0.664**	**0.875**	**0.546**	**0.935**	**0.532**	**0.871**

- **PanNuKe:** This dataset offers a broad diversity with 19 distinct tissue types for nuclei segmentation and classification. It includes 4,346 images for training and 1,888 images for testing, each measuring 256 × 256 pixels, showcasing a wide variety of tissues and nuclei types for comprehensive segmentation analysis.

7.5. Cross-modal Retrieval Datasets

Mirroring the approaches of PLIP and QuiltNet, we assessed the effectiveness of cross-modal retrieval through zero-shot text-to-image and image-to-text retrieval tasks, using the Twitter validation [35] and ARCH [28] datasets.

- **ARCH:** dataset, designed specifically for computational pathology, consists of 25,028 vision-language pairs from PubMed articles and pathology textbooks, narrowed down to 8,176 pairs after filtering, providing detailed diagnostic and morphological information across various stains, tissue types, and pathologies.
- **Twitter:** dataset, derived from pathology-related hashtags [4, 35, 53], offers a comprehensive collection of 243,375 public pathology images, including H&E stained cases, along with image-text pairs from tweets and replies. For our analysis, we utilized a validation subset similar to PLIP, comprising 2,023 image-text pairs filtered from over 200,000 pairs, to evaluate our model's cross-modal retrieval performance on both common and rare pathology cases.

8. More Ablation Studies

8.1. Impact of Magnification Levels (Table 1)

In Sec. 1, we presented the argument that multi-resolution pre-training is advantageous, considering that downstream CPath tasks are executed at various magnifications. Our experimental results indeed confirm that incorporating multiple resolutions in the pre-training dataset improves the performance of downstream tasks (see Table 1 for details).

To substantiate this finding, we carried out an ablation study by performing zero-shot classification tasks using combinations of different resolutions. We noted that the combination of 5×,10×,20×, and 40× resolutions yielded the highest performance. On average, using 10×, 20×, and 40× resolutions was the second most effective, outperforming the use of just 20× and 40×.

The rationale is that images and their paired textual descriptions at lower magnifications, like 5×, provide sufficient contextual information, whereas higher magnifications, like 40×, offer the necessary cellular detail for high performance. Current SOTA methods fail to capture either the broader context or the intricate details, leading to a decrease in performance.

8.2. Optimal Number of Positive Keywords (k_o) for the CVTA Module (Table 2)

We explore how different numbers of positive keywords (k_o), as detailed in Sec. (3.3), impact performance. According to the results shown in Table 2, enhancing k_o from 3 to 9 leads to performance gains across all datasets, attributed to the enrichment of information with more keywords. However, a further increment to $k_o = 12$ only marginally diminishes performance across most datasets, with the exception of DigestPath, which still shows improvement. Beyond this point, performance suffers due to the addition of noisy keywords that dilute the textual description's relevance.

8.3. Performance Comparison of Different Captioning Models (Table 3)

Table 3 presents the zero-shot classification performance of MR-PLIP across six datasets using four different captioning methods: Quilt-LLaVA [59], QuiltNet [37], BLIP2 [50], and GPT4V [2]. The best performance is observed with Quilt-LLaVA. Therefore, all results in this work are reported using Quilt-LLaVA as the captioning model.

8.4. Generalization to other Image-Text Encoders (Table 4)

In this experiment, we compared the performance of our proposed MR-PLIP model in terms of initializing different image-text encoders including CLIP (out-of-domain pre-trained encoders), PLIP (in-domain pre-trained encoders), CTransPath (in-domain pre-trained image encoder), PubMedBERT (out-of-domain pre-trained text encoders), BioclinicalBert (out-of-domain pre-trained text encoders), DinoSSLPath (in-domain pre-trained image encoder), QuiltNet (in-domain pre-trained encoders) as shown in Table

Table 2. Ablation study examining the zero-shot classification performance of the MR-PLIP model in terms of weighted F_1 score, while varying the number of positive keywords (k_0). The table highlights the effect of different k_0 values on the classification results across six datasets, with the optimal performance marked in bold for each dataset.

k_0 value	CAM16	CPTAC	SICAP	DigestPath	Databiox	NCT-CRC
$k_0 = 3$	0.602	0.816	0.514	0.918	0.476	0.813
$k_0 = 6$	0.636	0.858	0.529	0.926	0.484	0.847
$k_0 = 9$	**0.664**	**0.875**	**0.546**	0.935	**0.532**	**0.871**
$k_0 = 12$	0.662	0.859	0.532	**0.938**	0.498	0.863
$k_0 = 15$	0.642	0.824	0.528	0.929	0.481	0.859
$k_0 = 18$	0.611	0.821	0.505	0.901	0.470	0.850

Table 3. Ablation study comparing the zero-shot classification performance of MR-PLIP in terms of weighted F_1 score using different captioning models to generate textual descriptions of histology images at the multi-resolution level. The table highlights the performance of Quilt-LLaVA, QuiltNet, BLIP2, and GPT4V across six datasets, with the best performance for each dataset marked in bold.

Models	Quilt-LLaVA	QuiltNet	BLIP2	GPT4V
CAM16	**0.664**	0.621	0.577	0.634
CPTAC	**0.875**	0.834	0.788	0.817
SICAP	**0.546**	0.507	.468	0.433
DigestPath	**0.935**	0.881	0.821	0.846
Databiox	**0.532**	0.510	0.401	0.476
NCT-CRC	**0.871**	0.837	0.703	0.752

9. The best results on six datasets are reported using UNI (ViT-L/16-224) as an image encoder and QuiltNet (GPT-2/77) to initialize the text encoder. This is because UNI is pre-trained on unlabeled large histology images, and the QuiltNet text encoder is trained on 1M histology image-text pairs. The in-domain MR-PLIP variants also showed comparable performance compared to the best-performing MR-PLIP (in-domain) variant.

8.5. Impact of Zero-Shot Inference (Figs. 5-6 & Table 5)

In this experiment, we evaluate the performance of the MR-PLIP algorithm by employing two different zero-shot inference protocols. Echoing the SOTA approaches such as PLIP [35], QuiltNet [37], and MI-Zero [52], we extract visual features for a given histology patch using our vision encoder and compare them against a predefined set of testing prompts to determine its class label as shown in Fig. 5.

Since our VLM is based on fine-tuning two uni-modal encoders (one vision encoder and one text encoder) and one multi-modal encoder, we introduce a novel zero-shot approach (Fig. 6) where, for a specific test histology patch, we select k_o positive keywords from a dictionary of textual descriptions collected during the training process of the CVTA module (Sec. 3.3 in the main manuscript). These positive keywords and the visual features are input to the multi-modal encoder to obtain text-guided visual features which are then used to match with the testing prompts to predict the class label as shown in Fig. 6.

Table 5 shows the performance comparison of using both zero-shot inference protocols on six independent datasets in terms of weighted F_1 score. Following [37] and [52], we used similar testing prompts to compare the performances of both zero-shot evaluation protocols. Our proposed zero-shot inference protocol outperformed the classical approach across six datasets. This demonstrates that the learned text-guided visual representations using the proposed multi-modal encoder are more effective compared to the unimodal encoder representations.

8.6. Motivation of using Positive Keywords vs. Full Text (Table 6)

Synthetically generated textual descriptions may contain hallucinations, noise, and irrelevant words, which the proposed CVTA module (Sec. 3.2) removes. The top-k_o well-aligned words with visual features are retained as positive keywords. *By using positive keywords, we capture fine-grained tissue morphology in multi-resolution histology images, enhancing zero-shot classification*. In contrast, full-text descriptions may obscure such details. Moreover, aligning multi-resolution positive keyword representations with histology images enables MR-PLIP to localize meaningful tissue structures more effectively. Our positive keyword-based alignment approach improves generalization to novel keyword-tissue structures. *As shown in our ablation study (Table 6), positive keyword alignment outperforms full-text and all-word alignments across six datasets*.

9. Zero-Shot Experiments

Zero-shot learning refers to the capability of models to accurately perform tasks on new, unseen data without direct training on those specific tasks, utilizing pre-learned representations from image-text pairs.

Table 4. Zero-shot classification performance comparison in terms of weighted F_1 score using different pre-trained vision and text encoders.

Ablation Study	Vision Encoder	Text Encoder	CAM16	CPTAC	SICAP	DigestPath	Databiox	NCT-CRC
MR-PLIP	CLIP (ViT-B/16-224)	CLIP (GPT-2/77)	0.483	0.702	0.398	0.761	0.251	0.721
MR-PLIP	PLIP (ViT-B/32-224)	PLIP (GPT-2/347)	0.541	0.729	0.461	0.834	0.452	0.749
MR-PLIP	CTransPath (ViT-B/16-224)	BioClinicalBert (BioClinicalBert/512)	0.584	0.752	0.531	0.883	0.471	0.751
MR-PLIP	CTransPath (ViT-B/16-224)	PubMedBERT (PubMedBERT/256)	0.581	0.799	0.528	0.881	0.471	0.761
MR-PLIP	PLIP (ViT-B/32-224)	PubMedBERT (PubMedBERT/256)	0.556	0.765	0.500	0.841	0.440	0.769
MR-PLIP	PLIP (ViT-B/32-224)	BioClinicalBert (BioClinicalBert/512)	0.534	0.756	0.491	0.804	0.450	0.781
MR-PLIP	CTransPath (ViT-B/16-224)	PLIP (GPT/347)	0.563	0.771	0.503	0.924	0.479	0.806
MR-PLIP	QuiltNet (ViT-B/16-224)	QuiltNet (GPT-2/77)	0.591	0.773	0.498	0.871	0.451	0.774
MR-PLIP	QuiltNet (ViT-B/16-224)	PubMedBERT (PubMedBERT/256)	0.611	0.785	0.519	0.909	0.485	0.821
MR-PLIP	DinoSSLPath (ViT-B/16-224)	PubMedBERT (PubMedBERT/256)	0.634	0.802	0.538	0.914	0.486	0.835
MR-PLIP	DinoSSLPath (ViT-B/16-224)	QuiltNet (GPT-2/77)	0.622	0.841	0.541	0.931	0.491	0.842
MR-PLIP	DinoSSLPath (ViT-B/16-224)	BioClinicalBert (BioClinicalBert/512)	0.589	0.783	0.528	0.901	0.472	0.841
MR-PLIP	**UNI (ViT-L/16-224)**	**QuiltNet (GPT-2/77)**	**0.664**	**0.875**	**0.546**	**0.935**	**0.532**	**0.871**

Table 5. Zero-shot transfer for histology image classification performance comparison in terms of weighted average F_1 score. The results are reported using the zero-shot inference protocol used in the SOTA methods [35, 37, 52] and our proposed zero-shot transfer protocol.

Zero-Shot	CAM16	CPTAC	SICAP	DigestPath	Databiox	NCT-CRC
Classical Zero-shot	0.636	0.812	0.526	0.902	0.472	0.841
Proposed Zero-shot	**0.664**	**0.875**	**0.546**	**0.935**	**0.532**	**0.871**

Table 6. Zero-shot weighted F_1 scores for MR-PLIP using full text, all words, and positive keywords alignment.

Alignment	NCT-CRC	SICAP	Databiox	CAM16	CPTAC	EBRAINS
Full Text	0.834	0.446	0.481	0.611	0.833	0.351
All Words	0.846	0.458	0.501	0.629	0.846	0.376
Positive Keywords	**0.871**	**0.546**	**0.532**	**0.664**	**0.875**	**0.398**

9.1. Zero-shot Segmentation Results (Table 7)

We conduct zero-shot WSI-level segmentation using a process akin to the one for tile-based classification previously mentioned. Rather than compiling scores from tiles into a singular WSI-level prediction, we map tile-level scores back to their respective spatial locations within the WSI, averaging scores in overlapping areas. The highest-scoring class at each location is used to determine the pixel-level segmentation mask. Table 7 presents the zero-shot WSI-level segmentation outcomes on three datasets, setting them against six SOTA methods. Our MR-PLIP model outperforms all other methods in terms of performance across all three datasets. Overall, CPLIP secures its position as the runner-up in performance on the DigestPath and SICAP datasets, while QuiltNet ranks as the second-best on the TIGER dataset.

9.2. Zero-shot Cross-modal Retrieval Results (Table 8)

The zero-shot text-to-image and image-to-text retrieval tasks are evaluated by locating the closest matches for each modality and verifying whether the correct ground-truth pair falls within the top 1, 50, 200 closest matches. Table 8 shows the zero-shot cross-modal retrieval performance on two separate datasets, alongside a comparison with five SOTA VLMs in CPath. On both datasets, the MR-PLIP model outperforms all other methods by a significant margin, indicating its robust ability to align cross-resolution features across diverse textual and visual domains. CONCH and CPLIP also show strong performance, ranking as the second-best in terms of recall metrics.

Table 7. Zero-shot segmentation performance comparison of gigapixel images in terms of dice score, precision, and recall with existing VLMs in CPath on three independent datasets. The MR-PLIP algorithm outperforms existing models.

Methods	DigestPath[23]			SICAP [62]			TIGER [60]		
CLIP [57]	0.367	0.492	0.511	0.367	0.599	0.605	0.210	0.261	0.278
BioCLIP [73]	0.446	0.581	0.601	0.484	0.536	0.557	0.255	0.281	0.302
PLIP [35]	0.426	0.526	0.541	0.549	0.605	0.644	0.311	0.341	0.331
MI-Zero [52]	0.599	0.648	0.691	0.587	0.651	0.726	0.371	0.402	0.398
CONCH [51]	0.615	0.663	0.709	0.601	0.672	0.751	0.424	0.447	0.406
QuiltNet [37]	0.581	0.621	0.681	0.595	0.661	0.706	0.386	0.433	0.418
CPLIP [39]	0.687	0.722	0.761	0.651	0.715	0.806	0.420	0.454	0.413
MR-PLIP	**0.706**	**0.741**	**0.785**	**0.664**	**0.745**	**0.823**	**0.459**	**0.489**	**0.436**

Table 8. Zero-shot cross-modal retrieval (text-to-image and image-to-text) results on two datasets. In each cell, the results are displayed in the format (%|%), with text-to-image retrieval results on the left and image-to-text retrieval results on the right.

Methods	ARCH						Twitter					
	R@1		R@50		R@200		R@1		R@50		R@200	
CLIP	0.07	0.05	2.42	2.52	7.21	7.22	0.09	0.08	1.28	1.23	6.61	6.97
BioCLIP	8.89	9.97	53.24	52.13	71.43	68.47	9.11	10.56	40.30	39.23	52.33	51.66
PLIP	0.56	0.74	43.10	42.71	29.85	29.46	2.33	2.42	52.76	53.25	62.33	64.40
MI-Zero	6.87	7.71	52.10	54.14	60.96	61.21	5.77	6.98	70.21	68.83	75.66	74.10
QuiltNet	8.77	9.85	55.14	53.06	77.64	73.43	7.89	8.66	69.81	70.44	73.44	72.11
CONCH	8.16	9.11	58.91	59.10	75.16	76.90	9.90	10.19	75.91	76.80	80.91	81.68
CPLIP	9.10	9.06	56.77	57.12	79.10	80.19	6.17	7.90	72.89	73.09	79.10	80.17
MR-PLIP	**11.17**	**12.56**	**61.31**	**62.23**	**83.98**	**84.20**	**10.21**	**11.33**	**78.84**	**76.91**	**83.21**	**82.97**

Table 9. Performance comparison of the proposed MR-PLIP with existing SOTA foundation models, including both VLMs and vision-only models. The tile-level classification performance is reported using linear probe evaluations, while WSI-level classification results are reported using weakly supervised learning in which the ABMIL method is employed for both feature aggregation and MIL classification. BA represents balanced accuracy and F_1 is the weighted F_1 score.

Datasets	CONCH		QuiltNet		UNI		REMEDIS		Virchow		CHIEF		CTransPath		GigaPath		MR-PLIP	
(Tile-level)	BA	F_1	BA	F_1	BA	F_1	BA	F_1	BA	F_1	BA	F_1	BA	F_1	BA	F_1	BA	F_1
NCT-CRC	0.938	0.955	0.922	0.947	0.874	0.875	0.787	0.802	0.960	0.968	0.844	0.856	0.845	0.867	0.929	0.942	**0.965**	**0.976**
PatchCamelyon	0.866	0.869	0.822	0.831	0.901	0.930	0.805	0.822	0.933	0.933	0.833	0.851	0.911	0.935	0.925	0.931	**0.955**	**0.961**
WILDS-CAM17	0.911	0.925	0.861	0.877	**0.983**	**0.983**	0.926	0.926	0.971	0.971	0.901	0.922	0.960	0.960	0.951	0.962	0.975	0.980
MHIST	0.791	0.807	0.802	0.823	0.856	0.881	0.781	0.807	0.831	0.836	0.791	0.813	0.811	0.826	0.851	0.879	**0.876**	**0.915**
SICAP	0.711	0.745	0.722	0.767	0.826	0.841	0.806	0.811	0.855	0.873	0.771	0.783	0.678	0.747	0.845	0.861	**0.886**	**0.905**
WSSS4LUAD	0.811	0.825	0.805	0.812	0.831	0.835	0.769	0.782	0.866	0.873	0.812	0.828	0.844	0.857	0.860	0.872	**0.887**	**0.896**
BACH	0.856	0.871	0.833	0.861	0.925	0.926	0.863	0.864	0.915	0.920	0.847	0.863	0.875	0.872	0.933	0.947	**0.945**	**0.966**
UniToPatho	0.451	0.467	0.446	0.457	0.504	0.533	0.446	0.473	0.557	0.574	0.405	0.416	0.432	0.481	0.535	0.540	**0.605**	**0.622**
Datasets	CONCH		QuiltNet		UNI		REMEDIS		Virchow		CHIEF		CTransPath		GigaPath		MR-PLIP	
(WSI-level)	BA	F_1	BA	F_1	BA	F_1	BA	F_1	BA	F_1	BA	F_1	BA	F_1	BA	F_1	BA	F_1
CAM16	0.881	0.902	0.902	0.922	0.957	0.961	0.930	0.923	0.951	0.913	0.944	0.952	0.897	0.907	**0.967**	0.960	0.950	**0.966**
RCC-DHMC	0.856	0.866	0.851	0.864	0.919	0.926	0.865	0.877	0.922	0.931	0.897	0.901	0.804	0.883	0.921	0.936	**0.941**	**0.952**
HunCRC	0.681	0.721	0.702	0.722	0.643	**0.824**	0.604	0.787	0.621	0.667	0.651	0.667	0.556	0.728	0.641	0.667	**0.688**	0.701
BRCA-BRACS	0.723	0.748	0.718	0.725	0.687	0.691	0.676	0.696	0.708	0.722	0.656	0.666	0.639	0.648	0.704	0.715	**0.741**	**0.767**
PANDA	0.702	0.733	0.722	0.744	0.757	0.809	0.711	0.766	0.728	0.741	0.724	0.745	0.691	0.752	0.744	0.789	**0.786**	**0.816**
EBRAINS	0.687	0.717	0.655	0.666	0.675	0.746	0.382	0.471	0.701	0.723	0.688	0.706	0.514	0.597	0.687	0.704	**0.745**	**0.763**
NSCLC-CPTAC	0.881	0.902	0.877	0.900	0.904	0.935	0.841	0.866	0.923	0.936	0.922	0.934	0.877	0.895	0.900	0.915	**0.930**	**0.955**

10. Linear Probe Experiments (Table 9)

In the context of deep learning, linear probing refers to a technique used to evaluate the quality of features learned by a deep neural network. Specifically, it involves training a simple linear classifier (e.g., logistic regression) on top of the features extracted from a pre-trained neural network. This process is performed without fine-tuning the original network's weights; only the weights of the linear classifier are updated during the training process. The primary goal of linear probing is to assess how well the pre-trained network has captured valuable data representations. If a simple linear classifier can achieve high performance using the features extracted by the neural network, it suggests that the

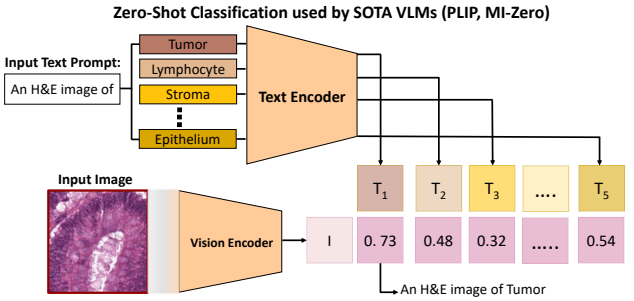


Figure 5. Illustration of the zero-shot classification process used by SOTA VLMs, like PLIP [35] and MI-Zero [52]. An input image, in this case, an H&E-stained tissue image indicating a tumor, is processed in parallel by a vision encoder and a text encoder. The vision encoder extracts features from the image, while the text encoder processes a set of predefined text prompts relating to possible classifications (e.g., Tumor, Lymphocyte, Stroma, Epithelium). Each class has a corresponding token $(T_1, T_2, T_3, ..., T_5)$, and the model outputs a probability score for each class, predicting the likelihood that the input image corresponds to each class based on visual-textual feature matching. The highest probability score indicates the model's classification of the image.

Table 10. WSI-level segmentation performance comparison in terms of dice score, precision, and recall of the MR-PLIP with other SOTA methods using **linear probe** evaluation protocols on three datasets.

Datasets	DigestPath			SICAP			TIGER		
PLIP	0.426	0.526	0.541	0.549	0.605	0.644	0.463	0.478	0.493
QuiltNet	0.521	0.545	0.564	0.592	0.603	0.621	0.602	0.619	0.627
DINOSSLPath	0.551	0.588	0.603	0.634	0.667	0.683	0.601	0.628	0.636
CTransPath	0.503	0.516	0.526	0.534	0.567	0.582	0.533	0.561	0.587
CONCH	0.615	0.663	0.709	0.601	0.672	**0.751**	0.433	0.457	0.461
UNI	0.804	0.811	0.826	0.645	0.662	0.603	0.687	0.702	0.724
Virchow	0.833	0.865	0.889	0.641	0.652	0.687	0.707	**0.732**	0.755
MR-PLIP	**0.851**	**0.873**	**0.903**	**0.655**	**0.688**	0.708	**0.726**	0.730	**0.778**

network has learned a rich and informative representation of the data.

Following the SOTA VLMs in CPath, we conducted a downstream analysis by freezing the weights of our proposed model and subsequently training linear layers for supervised classification tasks. We obtained text-guided visual features by inputting an image alongside its most closely matching text description from a predefined prompt set. A downstream linear classifier was then trained on these features for tile-level assessments to evaluate the quality of the representations learned by our MR-PLIP model.

Tables 9-11 displays the results of linear probe evaluations on three downstream histopathology tasks including tile-level classification, WSI-level segmentation, and nuclei segmentation across 21 datasets, evaluating the weighted average F_1 score and comparing it against SOTA CPath models including PLIP [35], MI-Zero [52], QuiltNet [37], CTransPath [68], and DINOSSLPath [41]. Across all datasets, our MR-PLIP algorithm consistently outperforms other methods by a substantial margin, affirming the advanced performance of the MR-PLIP model over the second-best performing methods GigaPath, Virchow, CONCH, and UNI.

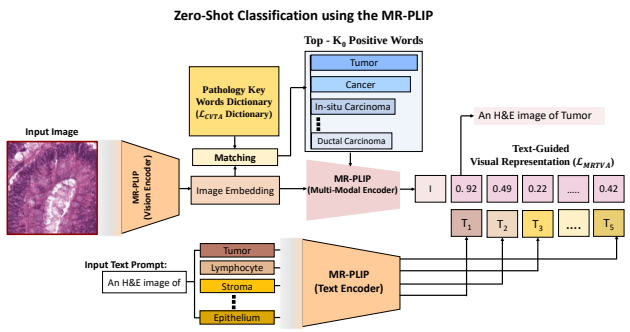


Figure 6. Illustration of the zero-shot classification process using the proposed MR-PLIP model. An input H&E (Hematoxylin and Eosin) stained tissue image is input to the vision encoder and a set of testing prompts are input to the text encoder. The vision encoder generates an image embedding, while the text encoder processes input text prompts such as "An H&E image of Tumor" and generates embedding. The image embeddings are matched against a pathology keyword dictionary, leading to a list of top k_0 positive words associated with various types of histology prompts. The multi-modal encoder (MR-PLIP) then combines these text and image embeddings to produce a text-guided visual representation. This yields a set of probability scores (such as 0.92, 0.49, etc.) for each possible classification token $T_1, T_2, T_3, ...T_5$, that represent different pathological features or diagnoses, predicting the most likely classification for the input image. The MR-PLIP model presents a significant enhancement through its integration of a pathology keywords dictionary, which augments the classification process. By comparing input text prompts with an extensive collection of relevant medical terminology provided by the dictionary, MR-PLIP can produce more precise and contextually detailed visual representations guided by text. As a result, it can determine more accurate classification probabilities for various pathology-specific tokens drawn from both the image and associated textual data. This approach advances beyond prior models by embedding specialized knowledge specific to the field of medicine into the algorithm's framework.

Table 11. Nuclei segmentation results in terms of mPQ measure of the MR-PLIP with other SOTA methods using both **linear probing** and **full fine-tuning** evaluation protocols on two datasets

Datasets	CONSEP		PanNuKe	
(Segmentation)	Linear	Fine-tune	Linear	Fine-tune
QuiltNet	41.11	45.56	48.11	52.21
MI-Zero	40.91	43.33	49.54	54.90
CONCH	39.85	51.75	44.31	50.21
DinoSSLPath	42.71	46.70	48.88	54.41
MR-PLIP	**46.61**	**52.25**	**51.64**	**58.61**

11. Weakly-Supervised WSI Classification Results (Table 9)

We performed weakly-supervised WSI classification to evaluate the text-guided visual representations learned by MR-PLIP across seven diverse WSI classification datasets. MR-PLIP was used to extract text-guided visual features from each patch, after which the ABMIL method [38] was employed for feature aggregation and MIL-based classification, as done in other SOTA methods [8, 19, 67–69]. For training, we used the AdamW optimizer with a cosine learning rate scheduler, a learning rate of 1×10^{-4}, cross-entropy loss, and a maximum of 20 epochs. To ensure fair comparisons, we followed the experimental protocols of existing SOTA methods for WSI classification tasks [19]. If official data folds were not available, the WSI datasets were case-stratified and label-stratified into train-validation-test splits as suggested by UNI [19].

Table 9 compares MR-PLIP with SOTA foundation models based on balanced accuracy and F_1 score. Our results show that MR-PLIP outperforms existing models by a significant margin, highlighting the benefits of explicitly incorporating multi-resolution image-text features.

12. Fine-tune Evaluation (Table 11)

In the context of deep learning, fine-tuning refers to a technique used to assess the adaptability and transfer potential of the learned weights by a deep neural network. Specifically, it involves fine-tuning the original network's weights. The primary goal of full fine-tuning is to assess how well the network weighted are transferred to the downstream analysis task.

We performed full fine-tuning of our model in conjunction with the linear layers for classification. This approach assesses the adaptability and transfer potential of the learned weights within the MR-PLIP framework.

Tables 11 display the results of full fine-tuning across nuclei segmentation datasets, evaluating the mPQ scores and comparing them against the SOTA methods. Across all datasets, our MR-PLIP algorithm consistently outperforms other methods by a substantial margin, affirming the advanced performance of the MR-PLIP model over the second-best performing methods QuiltNet, MI-Zero, and DINOSSLPath.

13. More Comparisons with SOTA MIL-based Methods (Table 12)

In this experiment, we compared the performance of our MR-PLIP model with recently proposed MIL-based methods including FiVE [47], R^{2}T [64], SI-MIL [42], ViLa-MIL [61], and PANTHER [63]. For a fair comparison, we employed the same ABMIL method for WSI-level feature aggregation and classification as discussed in Sec. 11.

Table 12 shows the results of the weakly supervised WSI classification and comparison with recently proposed SOTA methods on three tumor subtype classification datasets including CAM16, NSCL, and PANDA. The performance is reported in terms of balanced accuracy, F_1 score, and AUC. The MR-PLIP model consistently achieved superior performance across most evaluation metrics.

References

[1] Esther Abels, Liron Pantanowitz, Famke Aeffner, Mark D Zarella, Jeroen van der Laak, Marilyn M Bui, Venkata NP Vemuri, Anil V Parwani, Jeff Gibbs, Emmanuel Agosto-Arroyo, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *The Journal of pathology*, 249(3):286–294, 2019. 2

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 9

[3] Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412, 2014. 5

[4] Timothy Craig Allen. Social media: pathologists' force multiplier. *Archives of Pathology and Laboratory Medicine*, 138 (8):1000–1001, 2014. 9

[5] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019. 5

[6] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019. 3

[7] Harish Babu Arunachalam, Rashika Mishra, Ovidiu Daescu, Kevin Cederberg, Dinesh Rakheja, Anita Sengupta, David Leonard, Rami Hallac, and Patrick Leavey. Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. *PloS one*, 14(4):e0210706, 2019. 6

[8] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779, 2023. 14

[9] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018. 6, 7

[10] Carlo Alberto Barbano, Daniele Perlo, Enzo Tartaglione, Attilio Fiandrotti, Luca Bertero, Paola Cassoni, and Marco

Table 12. **Weakly-Supervised WSI Classification Results Compared with SOTA Methods**. Accuracy (Acc), weighted average F_1, and AUC scores for weakly-supervised WSI classification across multiple datasets. The proposed MR-PLIP model achieves the highest performance across most metrics, demonstrating its effectiveness over other SOTA methods.

Methods	CAM16			NSCLC			PANDA		
	Acc	F_1	AUC	Acc	F_1	AUC	Acc	F_1	AUC
PANTHER (CVPR24)[63]	0.933	0.949	<u>0.981</u>	<u>0.881</u>	0.867	0.906	0.751	0.796	<u>0.946</u>
R^2T (CVPR24)[64]	<u>0.954</u>	0.942	0.980	0.856	<u>0.883</u>	0.901	<u>0.768</u>	<u>0.806</u>	0.903
SI-MIL (CVPR24)[42]	0.944	<u>0.951</u>	0.963	0.844	0.830	0.881	0.733	0.771	0.886
ViLa-MIL (CVPR24)[61]	0.913	0.928	0.966	0.821	0.829	0.876	0.750	0.761	0.854
FiVE (CVPR24)[47]	0.942	0.932	0.975	0.842	0.853	<u>0.935</u>	0.743	0.788	0.873
MR-PLIP	**0.950**	**0.966**	**0.993**	**0.930**	**0.955**	**0.964**	**0.786**	**0.816**	**0.977**

Grangetto. Unitopatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 76–80. IEEE, 2021. 6

[11] Babak Ehteshami Bejnordi, Geert Litjens, Meyke Hermsen, Nico Karssemeijer, and Jeroen AWM van der Laak. A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In *Medical Imaging 2015: Digital Pathology*, pages 99–104. SPIE, 2015. 2

[12] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 7

[13] Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. Palm: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. *arXiv preprint arXiv:2004.07159*, 2020. 4

[14] Hamidreza Bolhasani, Elham Amjadi, Maryam Tabatabaeian, and Somayyeh Jafarali Jassbi. A histopathological image dataset for grading breast invasive ductal carcinomas. *Informatics in Medicine Unlocked*, 19:100341, 2020. 3, 6

[15] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022: baac093, 2022. 7

[16] Otso Brummer, Petri Pölönen, Satu Mustjoki, and Oscar Brück. Integrative analysis of histological textures and lymphocyte infiltration in renal cell carcinoma using deep learning. *bioRxiv*, pages 2022–08, 2022. 6

[17] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester Van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022. 7

[18] Chengkuan Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Andrew J Schaumberg, and Faisal Mahmood. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering*, 6(12):1420–1434, 2022. 2

[19] 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. 5, 8, 14

[20] X Chen, S Xie, and K He. An empirical study of training self-supervised vision transformers. in 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629. 4

[21] Toby C Cornish, Ryan E Swapp, and Keith J Kaplan. Whole-slide imaging: routine pathologic diagnosis. *Advances in anatomic pathology*, 19(3):152–159, 2012. 3

[22] Angel Cruz-Roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie NC Shih, John Tomaszewski, Fabio A González, and Anant Madabhushi. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific reports*, 7(1):46450, 2017. 3

[23] Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, et al. Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Medical Image Analysis*, 80:102485, 2022. 6, 12

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[25] Huijun Ding, Zhanpeng Pan, Qian Cen, Yang Li, and Shifeng Chen. Multi-scale fully convolutional network for gland segmentation using three-class classification. *Neurocomputing*, 380:150–161, 2020. 2

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[27] Nathan J Edwards, Mauricio Oberti, Ratna R Thangudu, Shuang Cai, Peter B McGarvey, Shine Jacob, Subha Madhavan, and Karen A Ketchum. The cptac data portal: a resource for cancer proteomics research. *Journal of proteome research*, 14(6):2707–2713, 2015. 7

[28] 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*, pages 16549–16559, 2021. 9

[29] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, pages 11–19. Springer, 2019. 8

[30] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019. 8

[31] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021. 5

[32] Chu Han, Xipeng Pan, Lixu Yan, Huan Lin, Bingbing Li, Su Yao, Shanshan Lv, Zhenwei Shi, Jinhai Mai, Jiatai Lin, et al. Wsss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. *arXiv preprint arXiv:2204.06455*, 2022. 6

[33] Matthew G Hanna, Anil Parwani, and Sahussapont Joseph Sirintrapun. Whole slide imaging: technology and applications. *Advances in Anatomic Pathology*, 27(4):251–259, 2020. 3

[34] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3852–3861, 2020. 2

[35] 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. 2, 3, 5, 6, 9, 10, 11, 12, 13

[36] BACH ICIAR. Grand challenge on breast cancer histology images. 2018, 2018. 6

[37] 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. 2, 3, 5, 6, 9, 10, 11, 12, 13

[38] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 14

[39] Sajid Javed, Arif Mahmood, Iyyakutti Iyappan Ganapathi, Fayaz Ali Dharejo, Naoufel Werghi, and Mohammed Bennamoun. Cplip: Zero-shot learning for histopathology with comprehensive vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11450–11459, 2024. 12

[40] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. 5

[41] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3344–3354, 2023. 6, 8, 13

[42] Saarthak Kapse, Pushpak Pati, Srijan Das, Jingwei Zhang, Chao Chen, Maria Vakalopoulou, Joel Saltz, Dimitris Samaras, Rajarsi R Gupta, and Prateek Prasanna. Si-mil: Taming deep mil for self-interpretability in gigapixel histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11226–11237, 2024. 14, 15

[43] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021. 6

[44] Katharina Kriegsmann, Frithjof Lobers, Christiane Zgorzelski, Joerg Kriegsmann, Charlotte Janssen, Rolf Ruedinger Meliss, Thomas Muley, Ulrich Sack, Georg Steinbuss, and Mark Kriegsmann. Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in Oncology*, 12:1022967, 2022. 6

[45] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 3, 4, 5

[46] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022. 4

[47] Hao Li, Ying Chen, Yifei Chen, Rongshan Yu, Wenxian Yang, Liansheng Wang, Bowen Ding, and Yuchen Han. Generalizable whole slide image classification with fine-grained visual-semantic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11398–11407, 2024. 14, 15

[48] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align be-

fore fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 3, 4, 5

[49] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 5

[50] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 9

[51] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, et al. Towards a visual-language foundation model for computational pathology. *arXiv preprint arXiv:2307.12914*, 2023. 3, 5, 6, 7, 12

[52] 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*, pages 19764–19775, 2023. 2, 3, 5, 6, 7, 8, 10, 11, 12, 13

[53] Michael J Misialek and Timothy Craig Allen. You're on social media! so now what? *Archives of Pathology & Laboratory Medicine*, 140(5):393–393, 2016. 9

[54] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979. 5

[55] Bálint Ármin Pataki, Alex Olar, Dezső Ribli, Adrián Pesti, Endre Kontsek, Benedek Gyöngyösi, Ágnes Bilecz, Tekla Kovács, Kristóf Attila Kovács, Zsófia Kramer, et al. Huncrc: annotated pathological slides to enhance deep learning applications in colorectal cancer screening. *Scientific Data*, 9(1):370, 2022. 7

[56] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 5

[57] 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*, pages 8748–8763. PMLR, 2021. 5, 12

[58] Thomas Roetzer-Pejrimovsky, Anna-Christina Moser, Baran Atli, Clemens Christian Vogel, Petra A Mercea, Romana Prihoda, Ellen Gelpi, Christine Haberler, Romana Höftberger, Johannes A Hainfellner, et al. The digital brain tumour atlas, an open histopathology resource. *Scientific Data*, 9(1):55, 2022. 7, 8

[59] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13183–13192, 2024. 1, 3, 9

[60] Adam Shephard, Mostafa Jahanifar, Ruoyu Wang, Muhammad Dawood, Simon Graham, Kastytis Sidlauskas, Syed Ali Khurram, Nasir Rajpoot, and Shan E Ahmed Raza. Tiager: Tumor-infiltrating lymphocyte scoring in breast cancer for the tiger challenge. *arXiv preprint arXiv:2206.11943*, 2022. 8, 12

[61] 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*, pages 11248–11258, 2024. 14, 15

[62] Julio Silva-Rodriguez, Adrián Colomer, Jose Dolz, and Valery Naranjo. Self-learning for weakly supervised gleason grading of local patterns. *IEEE journal of biomedical and health informatics*, 25(8):3094–3104, 2021. 6, 12

[63] Andrew H Song, Richard J Chen, Tong Ding, Drew FK Williamson, Guillaume Jaume, and Faisal Mahmood. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11566–11578, 2024. 14, 15

[64] 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*, pages 11343–11352, 2024. 14, 15

[65] Roetzer-Pejrimovsky Thomas, Moser Anna-Christina, Atli Baran, Clemens Christian Vogel, Petra A Mercea, Prihoda Romana, Ellen Gelpi, Haberler Christine, Höftberger Romana, Johannes A Hainfellner, et al. The digital brain tumour atlas, an open histopathology resource. *Scientific Data*, 9(1), 2022. 7, 8

[66] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018. 6

[67] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, pages 1–12, 2024. 14

[68] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022. 5, 13

[69] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, pages 1–9, 2024. 14

[70] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael

Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In *Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings*, pages 11–24. Springer, 2021. 6

[71] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013. 3, 5

[72] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15405–15416, 2023. 3

[73] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2024. 2, 3, 5, 12

[74] Zizhao Zhang, Pingjun Chen, Mason McGough, Fuyong Xing, Chunbao Wang, Marilyn Bui, Yuanpu Xie, Manish Sapkota, Lei Cui, Jasreman Dhillon, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1(5):236–245, 2019. 2

[75] Mengdan Zhu, Bing Ren, Ryland Richards, Matthew Suriawinata, Naofumi Tomita, and Saeed Hassanpour. Development and evaluation of a deep neural network for histologic classification of renal cell carcinoma on biopsy and surgical resection slides. *Scientific reports*, 11(1):7080, 2021. 7