CPLIP: Zero-Shot Learning for Histopathology with Comprehensive Vision-Language Alignment

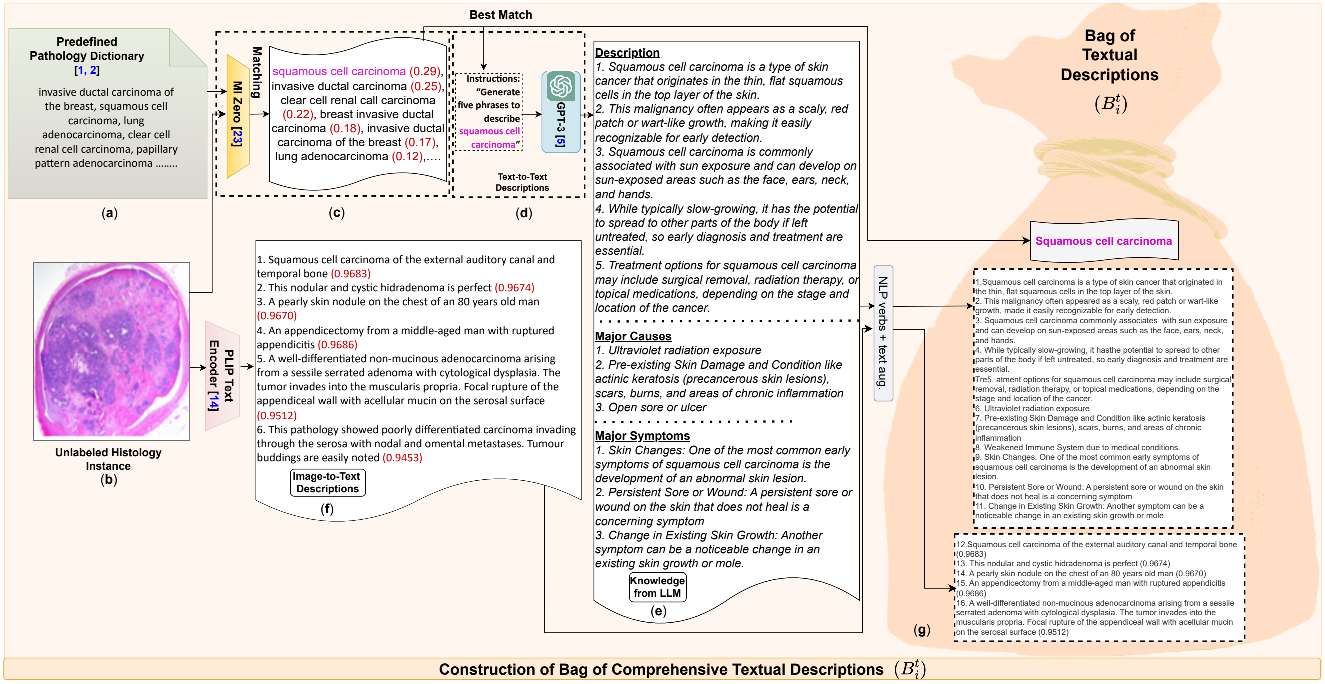
Sajid Javed[1], Arif Mahmood[2], Iyyakutti Iyappan Ganapathi[1,*], Fayaz Ali Dharejo[1],
Naoufel Werghi[1,*], Mohammed Bennamoun[3]
[1]Department of Computer Science, *C2PS, Khalifa University of Science and Technology, UAE
[2]Information Technology University of the Punjab, [3]The University of Western Australia

1. Glossary

- **An open vocabulary** allows machine learning models to recognize and work with words they haven't encountered before, rather than being limited to a pre-set list of terms.
- **Zero-shot transfer** refers to a machine learning model's ability to correctly handle tasks it has not been explicitly trained to perform, using knowledge learned during training from other related tasks.
- **Paired image and text data** consist of sets of images each directly associated with descriptive text that explains or provides context for the visual content. This pairing is used to train models to understand and align the content and context between the visual and textual information.
- **Whole Slide Images (WSIs)** are high-resolution digital scans of entire microscope slides containing tens of thousands of pixels used in pathology to examine tissues in detail.

  They are called **Whole Slide Images (WSIs)** because they are comprehensive digital scans that capture the entire tissue sample present on a glass slide, typically used for pathological examination. This allows pathologists to view the slide in its entirety on a computer, zoom in on areas of interest, and perform detailed analyses that would traditionally be done under a microscope.
- **Tile-level zero-shot learning** in the context of computational pathology refers to the ability of a machine learning model to classify individual tiles or patches of a whole slide image (WSI) into their correct categories without having been explicitly trained on those specific tiles or annotations. Each tile is a small, high-resolution section of a larger WSI, and the model must use learned patterns from other tasks or datasets to make accurate predictions.
- **Cancer subtyping** is the process of classifying cancer into more specific categories based on its cellular characteristics, molecular profile, and behavior. This helps in understanding the prognosis and determining the most effective treatment approach for each specific type.
- **Multi-Instance Learning (MIL)** is a variant of machine learning where data is grouped into 'bags' with a single label per bag, despite containing multiple instances. The MIL algorithm predicts bag labels by learning from the collective features of instances within each bag.
- **Is Multi-Instance Learning (MIL) considered to be a type of weakly supervised learning?** Yes, because it deals with training data that has incomplete or ambiguous labels. In MIL, only the bag of instances is labeled, not the individual instances, which is a weaker form of supervision than having labels for every instance.
- **MI-Zero** [19] is a framework designed to enhance the analysis of histopathology images, particularly gigapixel whole slide images used in medical diagnostics (see also main paper Sec. 3.2.1 for details). This framework is notable for its "zero-shot transfer capabilities." These capabilities are derived from contrastively aligned image and text models, which are used to facilitate multiple cancer subtype classification tasks.

  Specifically, MI-Zero uses pre-trained encoders to analyze these complex histopathology images. The key advantage of this approach is that it does not require any additional labeling of the images, which can be a time-consuming and resource-intensive process in medical image analysis. By leveraging existing models and their zero-shot transfer capabilities, MI-Zero aims to streamline and improve the diagnostic process in histopathology, enhancing the efficiency and accuracy of analyses conducted on these detailed images.
- **A lemma** is the base form of a word from which all its inflected or variant forms are derived. In the context of verbs, it's the form that appears in the dictionary, which is usually the present tense, singular form. For example, "go" is the lemma for "goes", "going", "went", and "gone". Lemmatization is the process of grouping together these different forms of a word so they can be analyzed as a single item. This is especially useful in natural language processing, where understanding the meaning of a word in different contexts is essential. **This type of augmentation is likely used to improve the model's un-**


Predefined Pathology Dictionary [1, 2]
MI Zero [23]
Matching
Best Match
Instructions: "Generate five phrases to describe squamous cell carcinoma"
GPT-3 [5]
Text-to-Text Descriptions
Unlabeled Histology Instance
PLIP Text Encoder [14]
Image-to-Text Descriptions
Knowledge from LLM
NLP verbs + text aug.
Bag of Textual Descriptions
(B_i^t)
Squamous cell carcinoma
(a)
(b)
(c)
(d)
(e)
(f)
(g)
Construction of Bag of Comprehensive Textual Descriptions (B_i^t)

invasive ductal carcinoma of the breast, squamous cell carcinoma, lung adenocarcinoma, clear cell renal cell carcinoma, papillary pattern adenocarcinoma ....

squamous cell carcinoma (0.29), invasive ductal carcinoma (0.25), clear cell renal call carcinoma (0.22), breast invasive ductal carcinoma (0.18), invasive ductal carcinoma of the breast (0.17), lung adenocarcinoma (0.12),....

1. Squamous cell carcinoma of the external auditory canal and temporal bone (0.9683)
2. This nodular and cystic hidradenoma is perfect (0.9674)
3. A pearly skin nodule on the chest of an 80 years old man (0.9670)
4. An appendicectomy from a middle-aged man with ruptured appendicitis (0.9686)
5. A well-differentiated non-mucinous adenocarcinoma arising from a sessile serrated adenoma with cytological dysplasia. The tumor invades into the muscularis propria. Focal rupture of the appendiceal wall with acellular mucin on the serosal surface (0.9512)
6. This pathology showed poorly differentiated carcinoma invading through the serosa with nodal and omental metastases. Tumour buddings are easily noted (0.9453)

Description
1. Squamous cell carcinoma is a type of skin cancer that originates in the thin, flat squamous cells in the top layer of the skin.
2. This malignancy often appears as a scaly, red patch or wart-like growth, making it easily recognizable for early detection.
3. Squamous cell carcinoma is commonly associated with sun exposure and can develop on sun-exposed areas such as the face, ears, neck, and hands.
4. While typically slow-growing, it has the potential to spread to other parts of the body if left untreated, so early diagnosis and treatment are essential.
5. Treatment options for squamous cell carcinoma may include surgical removal, radiation therapy, or topical medications, depending on the stage and location of the cancer.

Major Causes
1. Ultraviolet radiation exposure
2. Pre-existing Skin Damage and Condition like actinic keratosis (precancerous skin lesions), scars, burns, and areas of chronic inflammation
3. Open sore or ulcer

Major Symptoms
1. Skin Changes: One of the most common early symptoms of squamous cell carcinoma is the development of an abnormal skin lesion.
2. Persistent Sore or Wound: A persistent sore or wound on the skin that does not heal is a concerning symptom
3. Change in Existing Skin Growth: Another symptom can be a noticeable change in an existing skin growth or mole.

1.Squamous cell carcinoma is a type of skin cancer that originated in the thin, flat squamous cells in the top layer of the skin.
2. This malignancy often appeared as a scaly, red patch or wart-like growth, made it easily recognizable for early detection.
3. Squamous cell carcinoma commonly associates with sun exposure and can develop on sun-exposed areas such as the face, ears, neck, and hands.
4. While typically slow-growing, it hasthe potential to spread to other parts of the body if left untreated, so early diagnosis and treatment are essential.
Tre5. atment options for squamous cell carcinoma may include surgical removal, radiation therapy, or topical medications, depending on the stage and location of the cancer.
6. Ultraviolet radiation exposure
7. Pre-existing Skin Damage and Condition like actinic keratosis (precancerous skin lesions), scars, burns, and areas of chronic inflammation
8. Weakened Immune System due to medical conditions.
9. Skin Changes: One of the most common early symptoms of squamous cell carcinoma is the development of an abnormal skin lesion.
10. Persistent Sore or Wound: A persistent sore or wound on the skin that does not heal is a concerning symptom
11. Change in Existing Skin Growth: Another symptom can be a noticeable change in an existing skin growth or mole

12.Squamous cell carcinoma of the external auditory canal and temporal bone (0.9683)
13. This nodular and cystic hidradenoma is perfect (0.9674)
14. A pearly skin nodule on the chest of an 80 years old man (0.9670)
15. An appendicectomy from a middle-aged man with ruptured appendicitis (0.9686)
16. A well-differentiated non-mucinous adenocarcinoma arising from a sessile serrated adenoma with cytological dysplasia. The tumor invades into the muscularis propria. Focal rupture of the appendiceal wall with acellular mucin on the serosal surface (0.9512)

Figure 1. Diagram outlining the **detailed** construction process of the textual description bag (B_i^t) for best-matched prompt, "squamous cell carcinoma" shown in the main paper (Fig. 3 (A)). There are three primary steps: using MI-Zero to identify the best text match, leveraging GPT-3 to enrich the textual descriptions of the best-matched text, and employing the PLIP text encoder to generate more in-depth descriptions of the input unlabeled histology image.

derstanding by allowing it to recognize different forms of a verb as the same action or state.

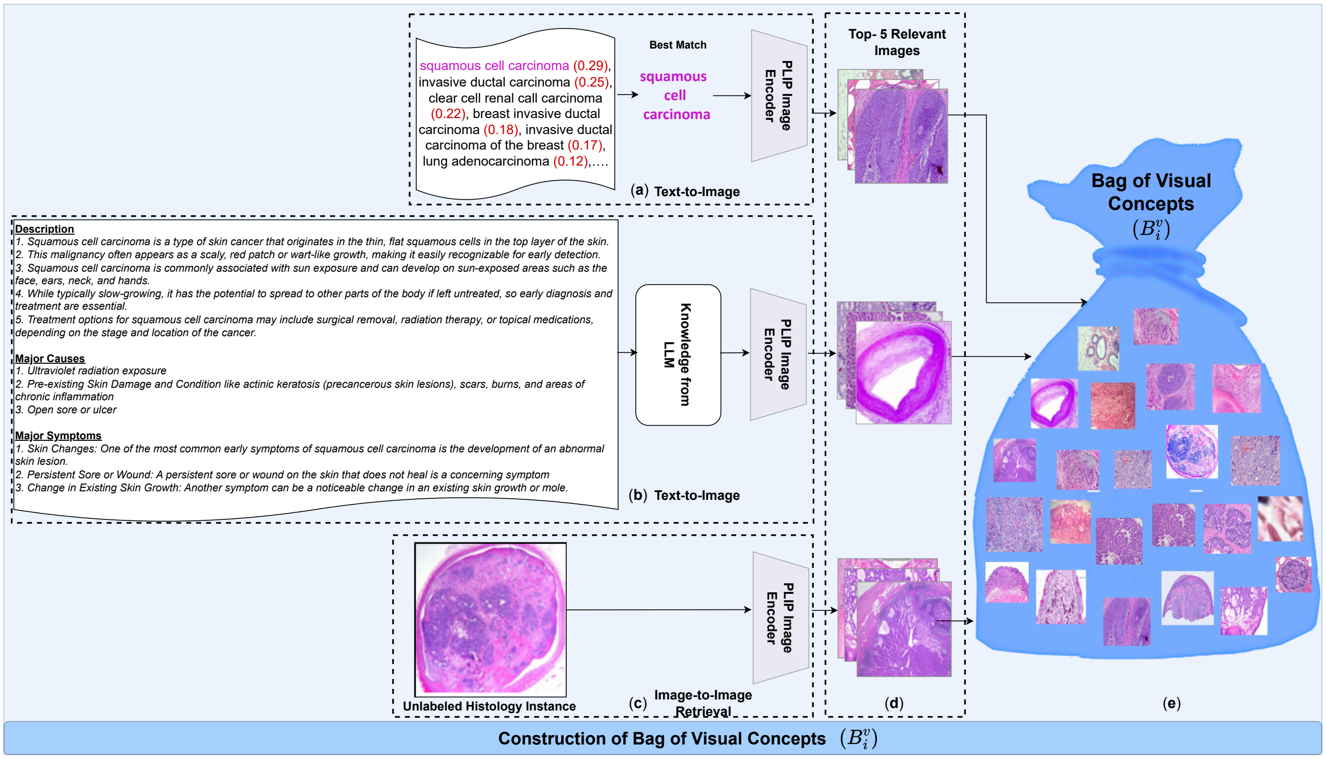
2. Our CPLIP Model Overview, Context, and Insights

In computational pathology, vision-language models have shown their impact in classifying and analyzing WSIs for various tasks (see PLIP [13], MI-Zero [19], BiomedCLIP [24], and CONCH [18]). The textual cues are instrumental in optimizing the performance of VL models. However, the current models' reliance on a singular prompt for a given histology image may lead to potentially restricted performance for zero-shot classification [13, 19, 24]. Typically, these models employ simple noun-based phrases like "`Photomicrograph showing clear cell change in oral squamous cell carcinoma`" or "`Photomicrograph of carcinomatous component (adenocarcinoma)`", overlooking the causes and symptoms associated with specific cancer types (please see main paper Fig. 2 (a) for details). Integrating more descriptive prompts, such as "`squamous cell carcinoma is instigated by exposure to ultraviolet radiation and human papillomavirus`" and "`symptoms of squamous cell carcinoma include skin changes, persistent sore or wound, or changes in existing skin growth`", could significantly enhance the information available to VL models during training (see Fig. 1).

To our knowledge, no existing computational pathology VL models have incorporated such diverse textual prompts either during training or at the inference stage. Unlike existing methods focusing on aligning individual textual and visual concepts, we propose a simultaneous alignment of numerous interrelated textual and visual concepts (refer to Fig. 2 (b) in the main paper for details).

We define "comprehensiveness" as the incorporation of a broad array of textual descriptions for the same medical conditions, coupled with a diverse set of histology images for those conditions. This approach acknowledges that a single disease may be described differently by various medical professionals and can manifest in multiple ways across patients. Despite these variances, combining different descriptions and images provides a holistic view, enhancing the VL models' ability to make connections between symptoms, causes, and specific medical conditions. As shown in Figs. 1 and 3, the best-matched prompt examples, "Squamous Cell Carcinoma" and "Sialdenoma papilliferum", have a broad array of textual descriptions coupled with a diverse set of causes and symptoms.

We exploit this **diversity** with a focus on "comprehen-



squamous cell carcinoma (0.29), invasive ductal carcinoma (0.25), clear cell renal call carcinoma (0.22), breast invasive ductal carcinoma (0.18), invasive ductal carcinoma of the breast (0.17), lung adenocarcinoma (0.12),....

Description
1. Squamous cell carcinoma is a type of skin cancer that originates in the thin, flat squamous cells in the top layer of the skin.
2. This malignancy often appears as a scaly, red patch or wart-like growth, making it easily recognizable for early detection.
3. Squamous cell carcinoma is commonly associated with sun exposure and can develop on sun-exposed areas such as the face, ears, neck, and hands.
4. While typically slow-growing, it has the potential to spread to other parts of the body if left untreated, so early diagnosis and treatment are essential.
5. Treatment options for squamous cell carcinoma may include surgical removal, radiation therapy, or topical medications, depending on the stage and location of the cancer.

Major Causes
1. Ultraviolet radiation exposure
2. Pre-existing Skin Damage and Condition like actinic keratosis (precancerous skin lesions), scars, burns, and areas of chronic inflammation
3. Open sore or ulcer

Major Symptoms
1. Skin Changes: One of the most common early symptoms of squamous cell carcinoma is the development of an abnormal skin lesion.
2. Persistent Sore or Wound: A persistent sore or wound on the skin that does not heal is a concerning symptom
3. Change in Existing Skin Growth: Another symptom can be a noticeable change in an existing skin growth or mole.

Figure 2. This diagram details the steps taken to create the bag of visual concepts B_i^v for the best-matched prompt "squamous cell carcinoma" shown in the main paper (Fig. 3 (B)). The process involves **(a)** using PLIP to select images that closely match the prompt, **(b)** using PLIP to enrich the dataset with histology images that align with the best-matched textual descriptions, and **(c)** employing PLIP to retrieve relevant histology images for the input unlabeled histology image

.

siveness". This term, in the context of textual prompts, refers to the variety of ways the same medical conditions (diseases) are described textually. Similarly, "comprehensiveness" in visual concepts involves having numerous histology images for the same medical condition. Our motivation is driven by the fact that medical practitioners often describe the same disease in various ways, and that the illness could manifest in varied forms for each patient. Despite these differences, the textual descriptions and the morphological characteristics of the disease are mutually informative. We propose to integrate detailed symptoms into the textual prompts, which would aid VL models in establishing correlations between symptoms, causes, and particular diseases or medical conditions. Moreover, we propose the integration of specific medical condition symptoms into the textual prompts, facilitating VL models in drawing correlations between symptoms, causes, and specific diseases or medical conditions.

To achieve comprehensive textual prompts, we initially compiled a dictionary of various cancer types and associated medical conditions by referencing multiple accessible online glossaries. Subsequently, for a given histology image, we assess the similarity and extract the most fitting prompts from the pathology dictionary using the existing VL model [19]. The best-matching prompt is then changed into five distinct variations using the GPT-3 model [5]. Additionally, we identified three primary causes and three notable symptoms related to the same prompt, utilizing GPT-3. Along with the same histology image, we also ascertain the most appropriate textual descriptions and relevant tissue images from the Medical Twitter dataset by using the PLIP model. Figs. 1 and 3 show two examples of best-matched prompts, "Squamous Cell Carcinoma" and "Sialdenoma papilliferum", with textual descriptions using GPT-3 and PLIP. To broaden the comprehensiveness/variety, we collect the most relevant tissue images through the PLIP model using the extended textual prompts obtained from GPT-3. Figs. 2 show the most relevant images associated with each textual description prompt analysed using the PLIP model. We restrict the count of textual descriptions and histology images to 17 and 21, although the number of these items could be increased based on the available computational resources.

From the comprehensive textual prompts and visual con-

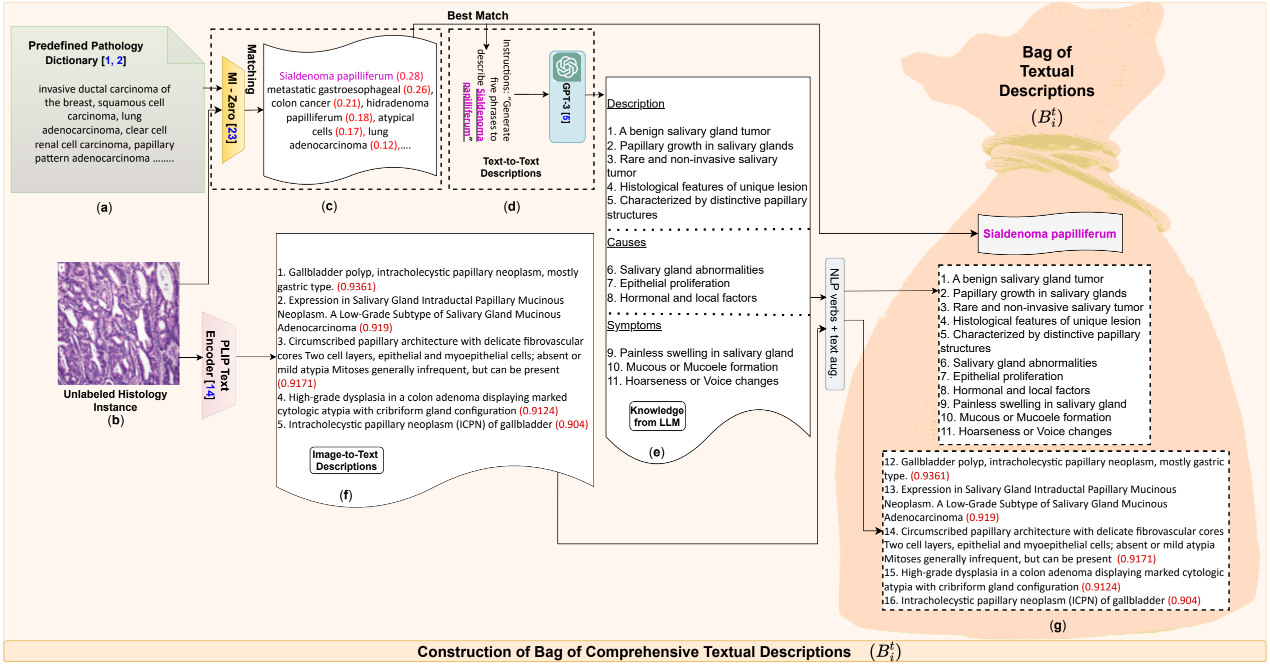

Best Match
Predefined Pathology Dictionary [1, 2]
invasive ductal carcinoma of the breast, squamous cell carcinoma, lung adenocarcinoma, clear cell renal cell carcinoma, papillary pattern adenocarcinoma ....
(a)
Matching
MI - Zero [23]
Sialdenoma papilliferum (0.28) metastatic gastroesophageal (0.26), colon cancer (0.21), hidradenoma papilliferum (0.18), atypical cells (0.17), lung adenocarcinoma (0.12),....
(c)
Instructions: "Generate five phrases to describe Sialdenoma papilliferum"
GPT-3 [5]
Text-to-Text Descriptions
(d)
Description
1. A benign salivary gland tumor
2. Papillary growth in salivary glands
3. Rare and non-invasive salivary tumor
4. Histological features of unique lesion
5. Characterized by distinctive papillary structures
Causes
6. Salivary gland abnormalities
7. Epithelial proliferation
8. Hormonal and local factors
Symptoms
9. Painless swelling in salivary gland
10. Mucous or Mucoele formation
11. Hoarseness or Voice changes
Knowledge from LLM
(e)
Unlabeled Histology Instance
(b)
PLIP Text Encoder [14]
1. Gallbladder polyp, intracholecystic papillary neoplasm, mostly gastric type. (0.9361)
2. Expression in Salivary Gland Intraductal Papillary Mucinous Neoplasm. A Low-Grade Subtype of Salivary Gland Mucinous Adenocarcinoma (0.919)
3. Circumscribed papillary architecture with delicate fibrovascular cores Two cell layers, epithelial and myoepithelial cells; absent or mild atypia Mitoses generally infrequent, but can be present (0.9171)
4. High-grade dysplasia in a colon adenoma displaying marked cytologic atypia with cribriform gland configuration (0.9124)
5. Intracholecystic papillary neoplasm (ICPN) of gallbladder (0.904)
Image-to-Text Descriptions
(f)
NLP verbs + text aug.
Bag of Textual Descriptions
(B_i^t)
Sialdenoma papilliferum
1. A benign salivary gland tumor
2. Papillary growth in salivary glands
3. Rare and non-invasive salivary tumor
4. Histological features of unique lesion
5. Characterized by distinctive papillary structures
6. Salivary gland abnormalities
7. Epithelial proliferation
8. Hormonal and local factors
9. Painless swelling in salivary gland
10. Mucous or Mucoele formation
11. Hoarseness or Voice changes
12. Gallbladder polyp, intracholecystic papillary neoplasm, mostly gastric type. (0.9361)
13. Expression in Salivary Gland Intraductal Papillary Mucinous Neoplasm. A Low-Grade Subtype of Salivary Gland Mucinous Adenocarcinoma (0.919)
14. Circumscribed papillary architecture with delicate fibrovascular cores Two cell layers, epithelial and myoepithelial cells; absent or mild atypia Mitoses generally infrequent, but can be present (0.9171)
15. High-grade dysplasia in a colon adenoma displaying marked cytologic atypia with cribriform gland configuration (0.9124)
16. Intracholecystic papillary neoplasm (ICPN) of gallbladder (0.904)
(g)
Construction of Bag of Comprehensive Textual Descriptions (B_i^t)

Figure 3. This figure outlines the process for creating B_i^t for the prompt "Sialdenoma papilliferum" as a second example. The procedure involves three primary steps: **(a)** using MI-Zero to select the best text match, **(b)** enriching the textual descriptions with GPT-3 to add depth, and **(c)** employing the PLIP text encoder to generate detailed descriptions of the input unlabeled histology image.

cepts, we generate a bag of textual descriptions and a bag of images in an unsupervised and automated manner see two examples in Figs. 1-2. Histology images corresponding to the same textual prompt in the established dictionary are classified as positive instances, while those corresponding to different prompts are labeled as negative instances. Using these comprehensive bags of textual and visual concepts, we fine-tune the baseline PLIP model to bring the embeddings of multiple positive textual and visual concepts closer together while distancing the embeddings of their negative counterparts. This process aims to boost class-agnostic representations (refer to the main paper, Fig. 1 (b)).

Our proposed fine-tuned model, termed the Comprehensive PLIP (CPLIP), can be employed in various downstream zero-shot classification tasks. Through our proposed methodology, we have made progress towards enhancing the alignment between textual and visual embeddings by incorporating inclusive textual descriptions, which include disease symptoms and causes alongside multiple visual concepts. By minimizing a comprehensive contrastive loss function, we facilitate the alignment of multiple textual and visual elements, a strategy that has shown significant effectiveness in our experiments. This paves the way for future research in this field.

Building on this foundation, our proposed CPLIP model has demonstrated superior performance in a range of computational pathology tasks, including tile-based classification, WSI-level classification for cancer subtyping, and histology image segmentation in zero-shot settings without any further fine-tuning Our model demonstrated improved patch-based classification performance compared to existing SOTA methods on diverse datasets such as CRC100K [16], WSSS4LUAD [12], DigestPath [7], and PanNuke [10]. In addition, the proposed model has also obtained better results than the existing SOTA methods on TCGA-BRCA, TCGA-RCC, and TCGA-NSCLC datasets.

3. Predefined Pathology Vocabulary

Table 10 shows our pathology prompts dictionary collected from online cancer glossaries [1, 2]. Our pathology dictionary covers diverse cancer types and morphologies across various tissue types and includes terms commonly used by expert pathologists to describe various cancer forms, related medical conditions, and their prognoses through histology images. This serves as a foundational prompt to match any input unlabelled histology images and extract comprehensive textual descriptions in subsequent phases.

4. Training and Implementation Details

In the field of histopathology, publicly accessible image-text paired datasets are scarce. The only publicly available image-text histology dataset is the ARCH dataset [9], which consists of 8,617 image-text pairs derived from 12,676 journal articles on clinical and research pathology. We fine-tuned our proposed CPLIP algorithm using the ARCH dataset, which contains histology images without using their corresponding textual descriptions. We expanded each image into approximately 21 images and 17 different textual descriptions, resulting in a total of 180,000 images and 146,000 textual descriptions. While the ARCH dataset was also used by MI-Zero [19] during training, they employed image-text paired data, whereas our algorithm requires unpaired many-to-many image-text alignment.

We fine-tuned our proposed CPLIP model by initializing weights using different image encoders and text encoders. For a performance comparison of our CPLIP model under different settings, refer to Table 2. The following different image-text encoders were employed during fine-tuning our CPLIP model:

1. Similar to the other SOTA methods [13, 19, 24], we fine-tuned the baseline CLIP [21] with a ViT-B/16-224 [8] as image encoder and a GPT-2/77 [20] as text encoder (Table 2).
2. Given that the baseline CLIP is trained on out-of-domain paired data, we also fine-tuned our CPLIP model using a pathology domain-specific pre-trained PLIP [13] with a PLIP-ViT-B/32-224 as image encoder and a GPT-2/347 as text encoder (Table 2).
3. Additionally, we fine-tuned the CPLIP model using BioClinicalBert/512 [3] and PubMedBERT/256 [11] as the text encoders and CTransPath/224 [23] as the image encoder, similar to MI-zero and BiomedCLIP [24] (Table 2). BioClinicalBert and PubMedBERT are medical-specific non-pathology text encoders trained on biomedical and clinical corpora such as PubMed abstracts and MIMIC [14], while CTransPath is trained using self-supervised representation learning on a total of 15.5 million unlabeled histology patches. Both of these encoders utilized ViT-B/16.
4. We fine-tuned our CPLIP algorithm using BioClinicalBert/512 as the text encoder and PLIP-ViT-B/32-224 as the in-domain image encoder (Table 2).
5. We also fine-tuned our CPLIP algorithm using CTransPath/224 as the in-domain image encoder and PLIP-GPT/347 as the in-domain text encoder (Table 2).

Across all visual-language pre-training variants, we trained our models using a temperature parameter of 0.02, the AdamW optimizer [17] with an initial learning rate of 5×10^{-6}, and a cosine decay scheduler. We trained our models with a batch size of 256 for 50 epochs. We set the filtering thresholds δ_t and δ_v to discard 10% of the data from each bag. After filtering, the bag of textual descriptions was reduced to 15 items and the bag of visual concepts was reduced to 19 items. Experiments are conducted using both single and merged prompts at inference time similar to [18] for fair comparison.

5. Datasets

1 .CRC100K [16]: is a colorectal cancer dataset containing 224 × 224 pixels tiles captured at 0.5 microns per pixel extracted from 50 patients. The dataset contains nine distinct tissue types including colorectal adenocarcinoma epithelium, normal colon mucosa, smooth muscle, lymphocytes, mucus, cancer-associated stroma, adipose, background, and debris. The official training (100K images) and testing (7,180 images) splits are provided. For zero-shot tile-based classification, we used the testing split without any fine-tuning.

2. WSSS4LUAD [12]: is a lung adenocarcinoma dataset containing tiles with almost 200 × 500 pixels. The dataset contains three distinct classes: tumor, tumor-associated stroma, and/or normal. Similar to PLIP, we performed binary classification of Tumor Vs. Normal. The training dataset contains 7063 images while the testing data consists of 3028 images (2015 Tumor, 1013 Normal). For zero-shot tile-based classification, we used the testing split without any fine-tuning.

3. SICAP [22]: is a prostate cancer dataset for Gleason pattern classification consisting of 512 × 512 pixels tiles extracted from 155 WSIs. The official training split consists of 9,959 images from 124 WSIs and the testing split consists of 2,122 images from 31 WSIs. The dataset contains four labels as the primary Gleason pattern (3, 4, or 5) or as non-cancerous (NC). We employed the official testing split for zero-shot classification experiments.

4. PanNuke [10]: is a more diverse nuclei segmentation and classification dataset consisting of 19 different tissue types. The training and testing splits consist of 4346 and 1888 images with 256 × 256 pixels. Similar to the PLIP, we evaluate the zero-shot classification performance of the proposed algorithm for Tumor vs. Normal Benign classes using the testing split.

5. DigestPath [7]: is a dataset of colonoscopy H & E tissue sections consisting of 660 images. Similar to PLIP, we performed tile-based zero-shot classification for Tumor Vs. Normal on the testing split containing 18814 images. For zero-shot segmentation, we employed the official 250 images from 93 patients for which pixel-level lesion annotation for colorectal cancer tissue is provided for testing.

6. CAM16 [4]: is a breast cancer dataset for lymph node metastasis detection using gigapixel WSIs. The total number of WSIs is 400 with only slide-level labels are provided. The official training split contains 270 WSIs and the testing split contains 130 testing WSIs. In training, the total number of normal WSIs is 159, and that containing tumor regions of breast cancer metastasis is 111. For zero-shot WSI-level classification, we used only the official testing split.

7. TCGA-BRCA[1] **:** is a TCGA dataset of invasive breast carcinoma containing two types of WSIs including Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC). The total number of WSIs is 1048 of which 837 are IDC and 211 are ILC. For zero-shot WSI-level classification, similar to CONCH, the test set consists of 75 WSIs from each class with no patient-level overlap between the training and testing splits.

8. TCGA-RCC[1]**:** is a TCGA dataset of renal cell carcinoma containing three types of WSIs including Clear Cell Renal Cell Carcinoma (CCRCC), Papillary Renal Cell Carcinoma (PRCC), and Chromophobe Renal Cell Carcinoma (CHRCC). The total number of WSIs is 922 of which 519 are CCRCC, 294 are PRCC, and 109 are CHRCC. For zero-shot WSI-level classification, similar to the CONCH, the test set consists of 75 WSIs from each of the three classes. There is no patient-level overlap between the training and testing splits.

9. TCGA-NSCLC[1]**:** is a TCGA dataset of Non-Small Cell Lung Cancer (NSCLC) containing two types of WSIs including LUng AaDenocarcinoma (LUAD) and LUng Squamous cell Carcinoma (LUSC) cases. The total number of WSIs is 1041 of which 529 are LUAD and 512 are LUSC. For zero-shot WSI-level classification, similar to CONCH, the test set consists of 75 WSIs from each of the two classes. There is no patient-level overlap between the training and testing splits.

6. Evaluation Metrics

We employed different evaluation metrics to evaluate the performance classification and segmentation tasks. For the classification task, we employed balanced accuracy, weighted F_1 score, and AUCROC. Balanced accuracy is defined as the macro average of the recall of each class. The weighted F_1 score is computed by taking the average of the F_1 score (the harmonic mean of precision and recall) of each class, weighted by the support of each class. In the binary case, AUCROC is the area under the receiver operating curve, which plots the true positive rate against the false positive rate as the classification threshold is varied. AUCROC is generalized to the multi-class case by averaging over the AUCROC of all pairwise combinations of classes. For the segmentation task, we report the Dice score, which is the same as the F_1 score, and the precision and recall of the positive class. The same set of evaluation metrics are also used by recent SOTA computational pathology VL models [13, 18].

7. Ablation Studies

1. Zero-shot performance comparison using single vs. merged prompts (Table 1). In this experiment, we compared the zero-shot classification performance of the proposed CPLIP algorithm using single prompts vs. merged prompts at the inference step (see Table 1). For a fair comparison with earlier works [13, 19, 24], we have used the same set of merged prompts as employed by CONCH [18]. On all five datasets for tile-based zero-shot classification, significant performance improvement is observed which is in line with the previous studies [13, 18].

Table 1. Ablation 1: Zero-shot classification performance comparison in terms of weighted average F_1 score using single vs. merged prompts. Significant performance improvement is observed using the merged prompts.95% Confidence Interval (CI) is included in parentheses.

Ablation Study	Single Prompts	Merged Prompts
CRC100K	0.681(0.663, 0.702)	**0.844**(**0.833**, **0.856**)
DigestPath	0.856(0.875, 0.889)	**0.903**(**0.891**, **0.915**)
SICAP	0.388(0.375, 0.395)	**0.511**(**0.498**, **0.526**)
WSSS4LUAD	0.791(0.784, 0.805)	**0.882**(**0.876**, **0.894**)
PanNuke	0.757(0.741, 0.763)	**0.811**(**0.799**, **0.827**)

2. Zero-shot performance comparison using different image-text encoders (Table 2). In this experiment, we compared the performance of our proposed CPLIP algorithm 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), BioclinicalBert and PubMedBERT (out-of-domain pre-trained text encoders) as shown in Table 2. The best results on five datasets are reported using CTransPath as an image encoder and BioClinicalBert to initialize the text encoder. This is because CTransPath is pre-trained on unlabeled larger histology images and BioClinicalBert is trained on 2M clinical notes in the MIMIC-III v1.4 database [14]. The in-domain CPLIP variants also showed comparable performance compared to the best-performing CPLIP (out-of-domain) variant.

3. Zero-shot performance comparison using different sizes of bags (Table 3). Experiments are also performed by varying the sizes of both bags. In the textual bag (B^t), rank-1 best-matching textual description with the input image, rank-5, rank-10, rank-15, and all 17 textual descriptions are considered. The corresponding visual bags (B^v) also contain 1, 5, 10, 15, and 21 images. As the bag sizes increase, continuous improvements in performance are observed until bag size 15, as shown in Table 3. A further increase has caused a slight decrease in performance due to noisy textual descriptions.

4. Many-to-Many Vs. One-to-One Contrastive Learn-

[1] portal.gdc.cancer.gov

Table 2. Ablation 2: Zero-shot classification performance comparison in terms of weighted average F_1 score using different pre-trained image and text encoders. 95% Confidence Interval (CI) is included in parentheses. All experiments use ViT-B/16 as the image encoder and PubMedBERT or BioClinicalBert to initialize the text encoder. Please note the performance is reported using merged prompts.

Ablation Study	Vision Encoder	Text Encoder	CRC100K	DigestPath	SICAP	WSSS4LUAD	PanNuke
CPLIP (Out-of-domain)	CLIP (ViT-B/16-224)	CLIP (GPT-2/77)	0.611 (0.588,0.634)	0.803 (0.794, 0.812)	0.344 (0.305,0.383)	0.765 (0.731,0.796)	0.708 (0.692,0.714)
CPLIP (In-domain)	PLIP (ViT-B/32-224)	PLIP (GPT/347)	0.828 (0.802,0.841)	0.886 (0.873, 0.804)	0.502 (0.491,0.511)	0.804 (0.791,0.815)	0.802 (0.793,0.814)
CPLIP **(Out-of-domain)**	**CTransPath** **(ViT-B/16-224)**	**BioClinicalBert** **(BioClinicalBert/512)**	**0.844** **(0.833,0.856)**	**0.903** **(0.891,0.915)**	**0.511** **(0.498,0.526)**	**0.882** **(0.876,0.894)**	**0.811** **(0.799,0.827)**
CPLIP (Out-of-domain)	CTransPath (ViT-B/16-224)	PubMedBERT (PubMedBERT/256)	0.838 (0.828,0.847)	0.894 (0.885,0.905)	0.508 (0.491,0.518)	0.866 (0.863,0.881)	0.807 (0.793,0.819)
CPLIP (Out-of-domain)	PLIP (ViT-B/32-224)	PubMedBERT (PubMedBERT/256)	0.825 (0.804,0.874)	0.881 (0.854,0.913)	0.482 (0.441,0.517)	0.841 (0.822,0.863)	0.782 (0.756,0.815)
CPLIP (Out-of-domain)	PLIP (ViT-B/32-224)	BioClinicalBert (BioClinicalBert/512)	0.828 (0.811,0.840)	0.891 (0.880,0.905)	0.494 (0.455,0.531)	0.871 (0.851,0.891)	0.798 (0.766,0.823)
CPLIP (In-domain)	CTransPath (ViT-B/16-224)	PLIP (GPT/347)	0.831 (0.821,0.843)	0.892 (0.881,0.835)	0.496 (0.471,0.512)	0.844 (0.822,0.867)	0.777 (0.761,0.786)

Table 3. Ablation 3: Zero-shot classification performance comparison in terms of weighted average F_1 score for varying size of text bag ($\delta_t = 100\%$). 95% Confidence Interval (CI) is included in parentheses. Please note the performance is reported using merged prompts.

Bag size (B^t)	1	5	10	15	17
CRC100K	0.766(0.751,0.788)	0.788(0.751,0.812)	0.811(0.803,0.827)	**0.844**(**0.833**,**0.856**)	0.841(0.821, 0.861)
DigestPath	0.856(0.825, 0.882)	0.881(0.852, 913)	0.901(0.871, 0.9410)	**0.903**(**0.891**, **0.915**)	0.902(0.895, 0.916)
WSSS4LUAD	0.841(0.832,0.856)	0.856(0.846,0.867)	0.875(0.861,0.889)	**0.882**(**0.876**,**0.894**)	0.881(0.856, 0.916)
SICAP	0.401(0.360,0.443)	0.433(0.413,0.466)	0.471(0.453,0.498)	**0.511**(**0.498**,**0.526**)	0.499(0.472, 0.521)
PanNuke	0.766(0.733,0.792)	0.781(0.752,0.812)	0.803(0.791,0.811)	**0.811**(**0.799**,**0.827**)	0.815(0.791, 0.827)

Table 4. Zero-shot classification performance, in terms of weighted average F_1 score, is compared between two contrastive learning approaches: one-to-one (CPLIP$_o$) and many-to-many (CLIP), using a single prompt. Significant improvements in performance are seen with the use of the proposed many-to-many contrastive learning method, as demonstrated across four datasets.

Datasets	One-to-One (CPLIP$_o$)	Many-to-Many (CLIP)
CRC100K	0.656	**0.681**
SICAP	0.341	**0.388**
TCGA-BRCA	0.732	**0.786**
TCGA-RCC	0.821	**0.855**

ing Approach (Table 4). The many-to-many learning approach has two main advantages: **(1)** it better reflects actual medical practice. Pathologists evaluate WSIs using not just visual morphology, but also patient symptoms and medical knowledge about disease causes. By integrating these multiple data sources, the approach allows for more comprehensive clinical integration compared to previous one-to-one methods; **(2)** the approach enhances visual representations through augmented slide image inputs, capturing phenotypic diversity and improving model generalization. This methodology constructs comprehensive textual descriptions and corresponding visual concepts to enable VLMs to handle the complexity of pathology images and text. The approach is akin to multi-task learning as learning joint representations across related tasks can promote generalization - analogous to how multi-task learning leads to more robust models. To assess its efficacy, an ablation study (Table 4) was conducted, revealing that this novel strategy significantly boosts zero-shot learning performance on four different datasets. Additionally, when compared to four SOTA vision-language models fine-tuned on the same datasets (Table 5), the many-to-many contrastive learning-based CPLIP model demonstrated superior accuracy, highlighting the robustness of this new technique.

5. Comparing CPLIP with in-domain SSL single modality CNNs and ViTs: Table 6 compares the classification performance of CPLIP against domain-specific SSL CNNs and ViTs, namely DinoSSLPath [15] and MoCo v2 [6]. The study spanned four datasets at both WSI and tile levels, using zero-shot, linear evaluation, and full fine-tuning

Table 5. The zero-shot classification performance of SOTA methods, evaluated using a single prompt in terms of the weighted average F_1 score on data generated by our proposed approach, shows significant improvements across all SOTA models. CPLIP stands out as the top performer.

Datasets	TCGA-NSCLC	TCGA-RCC	WSSS4LUAD	DigestPath
CLIP	0.488	0.291	0.541	0.137
BiomedCLIP	0.733	0.714	0.571	0.671
PLIP	0.744	0.751	0.761	0.842
MI-Zero	0.811	0.815	0.762	0.823
CPLIP	**0.835**	**0.855**	**0.791**	**0.856**

Table 6. Comparative classification performance in terms of the weighted average F_1 score using 3 methods: zero-shot learning, linear evaluation, and fine-tuning of the proposed CPLIP, which uses merged prompts vs. DinoSSLPath and MoCo v2 (with ResNet50). Significant performance improvements are observed in case of linear evaluation and full fine-tuning of CPLIP.

Datasets	Evaluation	DinoSSLPath	MoCo v2	CPLIP
CAM16 (WSI-level)	Zero-shot	×	×	**0.632**
	Linear	0.618	0.592	**0.663**
	Fine-tune	0.722	0.678	**0.746**
WSSS4LUAD (Tile-based)	Zero-shot	×	×	**0.882**
	Linear	0.878	0.881	**0.924**
	Fine-tune	0.951	0.944	**0.976**
CRC100K (Tile-based)	Zero-shot	×	×	**0.844**
	Linear	0.862	0.853	**0.894**
	Fine-tune	0.945	0.911	**0.964**
SICAP (Tile-based)	Zero-shot	×	×	**0.511**
	Linear	0.502	0.466	**0.554**
	Fine-tune	0.604	0.547	**0.626**

evaluation protocols. Notably, zero-shot evaluations using CPLIP with merged prompts surpassed the performance of DinoSSLPath and MoCo v2, which did not employ zero-shot settings. In linear and fine-tuning settings, CPLIP achieved significant performance gains, maintaining consistency with the protocols established by DinoSSLPath.

8. Tile-level Zero-shot Classification Results (Table 7)

We performed zero-shot experiments on five different datasets for tile-level classification using only the testing split of each dataset. Table 7 shows the comparison with existing SOTA VL-based methods in terms of balanced accuracy, weighted average F_1, and AUROC scores using both single and merged prompts. Our proposed CPLIP algorithm achieved the best performance on all three metrics on all five datasets in both settings. CONCH obtained the second-best performance on the CRC100K and SICAP datasets, but its performance has not been reported on the remaining datasets.

Specifically, CPLIP obtained a 13.5% improvement in balanced accuracy, a 13.90% improvement in weighted F_1, and a 2.10% improvement in AUROC over CONCH on the CRC100K dataset using single prompts. Using merged prompts, CPLIP achieved a 3.20% improvement in balanced accuracy, a 4.10% improvement in weighted F_1, and a 0.10% improvement in AUROC over CONCH on the CRC100K dataset. On the SICAP dataset, CPLIP achieved a 1.70% improvement in balanced accuracy, and a 14.30% improvement in weighted F_1 over CONCH using single prompts. Using merged prompts, CPLIP achieved an 8.70% improvement in weighted F_1 over CONCH. For DigestPath and PanNuke datasets, CPLIP improved (0.6 %, 3.2%, 3.0%) and (5.10%, 5.20%, 3.50%) performance using merged prompts over the second best performers MI-Zero/PLIP. On the WSSS4LUAD dataset, CPLIP improved its performance over the second-best performer PLIP by 10.0% in balanced accuracy, 9.10% in weighted F_1, and 7.0% in AUROC using merged prompts. *Most of these performances are significantly better than the existing SOTA methods, demonstrating the advantages of our proposed CPLIP algorithm.*

8.1. WSI-level Zero-shot Classification Results (Table 8)

To extend zero-shot transfer to gigapixel WSIs, we used a method similar to MI-Zero [19]. For classification over C classes, we first used the OTSU method to binarize the WSI into tissue and background regions. We then divided the tissue region into N tiles, each of size 224×224 pixels. For each tile, we estimated an ℓ_2-normalized embedding independently using the CPLIP image encoder. For each tile embedding, we computed cosine similarity scores with each text embedding, obtaining a set of C similarity scores for each tile. To aggregate similarity scores across tiles, we used the top-K pooling operator, averaging over the highest K similarity scores for each class to obtain the slide-level similarity score. The class with the highest slide-level score was the predicted class. We chose $K \in 1, 5, 10, 50, 100$ and reported the results for the K with the highest balanced accuracy, weighted F_1, and AUROC scores for classification tasks.

In Table 8, we compared the zero-shot classification performance of our proposed CPLIP algorithm with existing SOTA VL-based computational pathology models on four independent datasets: CAM16, TCGA-BRCA, TCGA-RCC, and TCGA-NSCLC, using both single and merged prompts. We also presented the performance of our proposed CPLIP algorithm in terms of different fine-tuned out-of-domain and in-domain image and text encoders.

CPLIP outperformed SOTA in-domain VL models, including PLIP, BiomedCLIP, MI-Zero, and CONCH, on all datasets, often by a significant margin. For example, in the case of lymph node metastasis classification in

Table 7. Tile-level zero-shot classification performance comparison in terms of balanced accuracy, weighted F_1, and AUCROC scores with existing VL-based models in computational pathology on five independent external datasets. On the WSSS4LUAD dataset, CONCH used a different split for performance evaluation which is indicated by *. CPLIP performance is reported using the best combination from ablation study 2 (Table. 2).

Single Prompt	CRC100K	DigestPath	SICAP	WSSS4LUAD	PanNuke
CLIP baseline [21]	0.234|0.185|0.727	0.11|0.030|0.203	0.231|0.139|0.201	0.451|0.481|0.705	0.322|0.352|0.683
BiomedCLIP [24]	0.422|0.372|0.859	0.591|0.622|0.781	**0.381**|0.361|0.506	0.466|0.495|0.698	0.522|0.572|0.711
PLIP [13]	0.520|0.517|0.879	0.815|0.832|0.901	0.319|0.255|0.603	0.702|0.734|0.822	0.629|0.656|0.805
MI-Zero [19]	0.544|0.536|0.872	0.822|0.811|0.911	0.308|0.251|0.605	0.722|0.742|0.805	0.659|0.688|0.755
CONCH [18]	0.566|0.542|0.901	-	0.349|0.245|-	0.598*|0.590*|0.795*	-
Proposed CPLIP	**0.701|0.681|0.922**	**0.835|0.856|0.933**	0.366|**0.388|0.711**	**0.778|0.791|0.836**	**0.681|0.757|0.835**
Merged Prompts	CRC100K	DigestPath	SICAP	WSSS4LUAD	PanNuke
CLIP baseline [21]	0.271|0.247|0.781	0.188|0.210|0.280	0.283|0.191|0.205	0.501|0.544|0.791	0.385|0.412|0.744
BiomedCLIP [24]	0.553|0.533|0.924	0.644|0.671|0.831	0.483|0.439|0.605	0.511|0.533|0.764	0.631|0.651|0.802
PLIP [13]	0.674|0.687|0.944	0.865|0.871|0.931	0.355|0.315|0.656	0.751|0.791|0.833	0.719|0.744|0.874
MI-Zero [19]	0.721|0.755|0.956	0.844|0.866|0.941	0.341|0.306|0.641	0.741|0.781|0.846	0.744|0.759|0.901
CONCH [18]	0.791|0.803|0.979	-	**0.624**|0.424|-	0.719*|0.705*|0.877*	-
Proposed CPLIP	**0.823|0.844|0.980**	**0.871|0.903|0.971**	0.498|**0.511|0.716**	**0.851|0.882|0.903**	**0.795|0.811|0.936**

Table 8. WSI-level zero-shot classification performance comparison in terms of balanced accuracy, weighted F_1, and AUCROC scores with existing VL-based models in computational pathology on five independent external datasets. On the WSSS4LUAD dataset, CONCH used a different split for performance evaluation which is indicated by *. We employed similar merged prompts during inference as proposed in CONCH [18]. (OoD: Out-of-Domain, InD: In-Domain)

Models (Single prompts)	Image encoder pretraining	Text encoder pretraining	CAM16	TCGA-BRCA	TCGA-RCC	TCGA-NSCLC
CLIP baseline [21]	ViT-B/16-224	GPT-2/77	0.134|0.175|0.325	0.512|0.328|0.551	0.321|0.178|0.578	0.496|0.358|0.536
BiomedCLIP [24]	ViT-B/16-224	PMB/256	0.311|0.377|0.545	0.527|0.422|0.761	0.677|0.646|0.872	0.699|0.684|0.851
PLIP [13]	ViT-B/32-224	GPT/347	0.399|0.416|0.681	0.451|0.331|0.611	0.726|0.739|0.915	0.676|0.666|0.781
MI-Zero [19]	CTransPath/224	BioClinicalBert/512	0.456|0.461|0.755	0.781|0.723|0.856	0.805|0.782|0.881	0.802|0.792|0.866
CONCH [18]	ViT-B/16-256	HistPathGPT/512	-	0.643|0.600|0.873	0.796|0.797|**0.961**	0.807|0.803|0.915
CPLIP$_1$ (Ours)	ViT-B/16-224 (OoD)	GPT-2/77 (OoD)	0.502|0.477|0.705	0.500|0.544|0.722	0.754|0.749|0.865	0.761|0.788|0.821
CPLIP$_2$ (Ours)	PLIP-ViT-B/32-224 (InD)	PLIP-GPT/347 (InD)	**0.591|0.587|0.827**	**0.824|0.786|0.889**	**0.844|0.855**|0.926	**0.854|0.835|0.936**
Models (Merged Prompts)	Image encoder pretraining	Text encoder pretraining	CAM16	TCGA-BRCA	TCGA-RCC	TCGA-NSCLC
CLIP baseline [21]	ViT-B/16-224	GPT-2/77	0.151|0.198|0.331	0.534|0.346|0.623	0.367|0.219|0.651	0.567|0.431|0.598
BiomedCLIP [24]	ViT-B/16-224	PMB/256	0.337|0.402|0.564	0.532|0.441|0.837	0.807|0.773|0.903	0.777|0.761|0.861
PLIP [13]	ViT-B/32-224	GPT/347	0.446|0.442|0.711	0.487|0.364|0.655	0.794|0.772|0.935	0.768|0.805|0.819
MI-Zero [19]	CTransPath/224	BioClinicalBert/512	0.499|0.521|0.821	0.833|0.821|0.905	0.871|0.855|0.933	0.881|0.871|0.944
CONCH [18]	ViT-B/16-256	HistPathGPT/512	-	0.840|0.839|0.932	0.893|0.895|0.973	0.900|0.900|0.964
CPLIP$_1$ (Ours)	ViT-B/16-224 (OoD)	GPT-2/77 (OoD)	0.578|0.551|0.751	0.557|0.588|0.783	0.834|0.805|0.921	0.811|0.856|0.833
CPLIP$_2$ (Ours)	PLIP-ViT-B/32-224 (InD)	PLIP-GPT/347 (InD)	**0.661|0.632|0.886**	**0.887|0.871|0.963**	**0.941|0.937|0.978**	**0.931|0.951|0.981**

CAM16 using single and merged prompts, CPLIP$_2$ (in-domain) achieved zero-shot balanced accuracies of 59.10% and 66.10%, respectively, and outperformed the next best performing model, MI-Zero, by 13.50% and 16.20%.

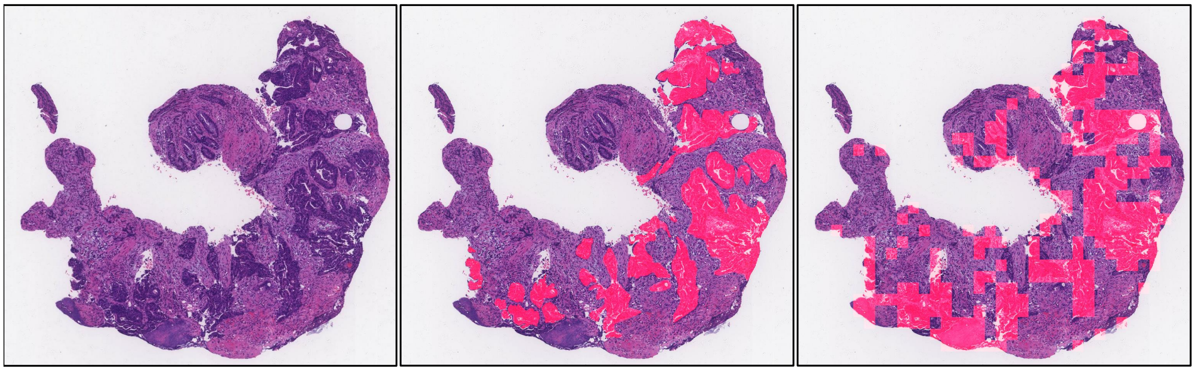
For Non-Small Cell Lung Cancer (NSCLC) and Renal Cell Carcinoma (RCC) subtyping using a single prompt, our proposed CPLIP$_2$ (In-domain) model achieves zero-shot balanced accuracies of 85.40% and 84.40% respectively. This outperforms the next best models, CONCH and MI-Zero, by margins of 4.70% and 5.20% on NSCLC and 4.80% and 3.90% on RCC. With merged prompts, CPLIP$_2$ (In-domain) further improves to 93.10% and 94.10% balanced accuracy, exceeding CONCH by 3.10% and 4.80%. Similarly, on the more challenging invasive breast carcinoma (BRCA) subtyping task, our CPLIP$_2$ (In-domain) achieves 88.70% zero-shot balanced accuracy, surpassing CONCH and MI-Zero by significant margins of 4.70% and 5.40%. *Overall, the proposed CPLIP$_2$ demonstrates SOTA performance on multiple cancer subtyping tasks using zero-shot learning.*

8.2. Zero-shot Segmentation Results of Gigapixel Images (Table 9)

We perform zero-shot segmentation of gigapixel WSIs similar to CONCH [18] using the same classification methods described above. We divide the WSI into tiles and com-

Table 9. Zero-shot segmentation performance comparison of gigapixel images in terms of dice score, precision, and recall with existing VL-based models in computational pathology on two independent datasets using the single prompt. OoD: Out-of-Domain and InD: In-Domain

Models (Single prompts)	Image encoder pretraining	Text encoder pretraining	SICAP	DigestPath
CLIP baseline [21]	ViT-B/16-224	GPT-2/77	0.367|0.599|0.605	0.367|0.492|0.511
BiomedCLIP [24]	ViT-B/16-224	PMB/256	0.484|0.536|0.557	0.446|0.581|0.601
PLIP [13]	ViT-B/32-224	GPT/347	0.549|0.605|0.644	0.426|0.526|0.541
MI-Zero [19]	CTransPath/224	BioClinicalBert/512	0.587|0.651|0.726	0.599|0.648|0.691
CONCH [18]	ViT-B/16-256	HistPathGPT/512	0.601|0.672|0.751	0.615|0.663|0.709
CPLIP$_1$ (Ours)	ViT-B/16-224 (OoD)	GPT-2/77 (OoD)	0.591|0.661|0.681	0.491|0.581|0.602
CPLIP$_2$ (Ours)	PLIP-ViT-B/32-224 (InD)	PLIP-GPT/347 (InD)	**0.654**|0.704|0.803	0.685|0.719|0.754
CPLIP$_3$ (Ours)	CTransPath/224 (InD)	BioClinicalBert/512 (OoD)	0.633|0.702|0.791	0.665|0.711|0.744
CPLIP$_4$ (Ours)	CTransPath/224 (InD)	PLIP-GPT/347 (InD)	0.651|**0.715**|**0.806**	**0.687**|**0.722**|**0.761**


(a) Whole Slide Image

(b) GroundTruth

(c) Zero-shot Prediction using CPLIP

Figure 4. Example of the segmentation results on one of the WSIs selected from DigestPath [7] dataset using our proposed CPLIP algorithm. Here, it is important to note that the segmentation task is posed as a tile-based zero-shot classification problem similar to CONCH [18]. The WSI is divided into tiles and the similarity scores for each tile are computed independently. However, instead of aggregating the scores across tiles into a single slide-level prediction, we map the tile-level scores to their corresponding spatial locations in the WSI and average the scores in the overlapping regions. Finally, each pixel is labeled with the class that has the top score, resulting in the formation of a detailed pixel-wise segmentation mask.

pute similarity scores for each tile independently. However, instead of aggregating the scores across tiles into a single slide-level prediction, we map the tile-level scores to their corresponding spatial locations in the WSI and average the scores in overlapped regions. Finally, for each pixel, we assign the class with the highest score as the prediction, producing a pixel-level segmentation mask.

We used the official testing splits of the SICAP dataset (31 WSIs) for prostate tumor vs. normal tissue segmentation and DigestPath (250 large images) for colon malignant vs. benign tissue for zero-shot segmentation. Results are reported in Table 9 in terms of Dice score, precision, and recall to quantify the quality of the predicted segmentation mask relative to the ground truth using a single prompt during the inference stage. Our proposed CPLIP algorithm outperforms other VL computational pathology models in both datasets. A visual result of the CPLIP algorithm is shown in Fig. 4 using a sample WSI from the DigestPath dataset.

In SICAP, our best-performing CPLIP$_2$ and CPLIP$_4$ models achieve average Dice scores of 65.40% and 65.10%, respectively, outperforming CONCH (60.10%), MI-Zero (58.70%), PLIP (54.90%), and BiomedCLIP (48.40%) by a significant margin. In DigestPath, our proposed best-performing in-domain CPLIP$_4$ and CPLIP$_2$ models achieved average Dice scores of 68.70% and 68.50%, respectively, outperforming CONCH (61.50%), MI-Zero (59.90%), PLIP (42.60%), and BiomedCLIP (44.60%) by a significant margin. *Additionally, we found that despite the coarse-grained and zero-shot nature of the approach, CPLIP was able to produce reasonably accurate pixel-level segmentation masks, demonstrating the advantages of heterogeneous textual descriptions and histology images.*

8.3. Computational Time Analysis

We conducted our experiments on a DGX NVIDIA workstation with 256 GB of RAM and 4 Tesla V100 GPUs. At inference, our model only needs to first compute the image-text representation and then perform cosine similarity, which can be implemented efficiently using matrix multiplication. On the TCGA-BRCA dataset for cancer subtyping, CPLIP took an average of 3.2 minutes to process per WSI, depending on the value of K, while other VL models, including PLIP (2.90 minutes), MI-Zero (3.00 minutes), and BiomedCLIP (2.70 minutes), were faster. Overall, CPLIP is comparable in speed to other VL models at inference.

References

[1] https://lab-ally.com/histopathology-resources/histopathology-glossary/. . 4

[2] https://www.cancer.org/cancer/understanding-cancer/glossary.html. . 4

[3] 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

[4] 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. 5

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020. 3

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020. 7

[7] 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. 4, 5, 10

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021. 5

[9] 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. 4

[10] 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. 4, 5

[11] 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

[12] 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. 4, 5

[13] 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, pages 1–10, 2023. 2, 5, 6, 9, 10

[14] 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, 6

[15] 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. 7

[16] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine, 16(1):e1002730, 2019. 4, 5

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019. 5

[18] 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. 2, 5, 6, 9, 10

[19] 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. 1, 2, 3, 5, 6, 8, 9, 10

[20] 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

[21] 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, 9, 10

[22] 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. 5

[23] 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

[24] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915, 2023. 2, 5, 6, 9, 10

Table 10. Our proposed pathology prompts dictionary used during the construction of a bag of textual descriptions and a bag of visual concepts.

Alphabet	Pathology Prompts Dictionary
A	Advanced breast Cancer; Antibody-Dependent cellular cytotoxicity; Adenocarcinoma; Adenoma benign cancer;
	Adenomatous polyp; Adenocarcinoma of the lung; Atypical glandular cells; Acinar pattern
	adenocarcinoma; Acinar growth pattern; Acinar predominant histological subtype;
	Alanine aminotransferase / alanine transaminase; Anaplastic Large-cell Lymphoma; Acute lymphocytic leukemia;
	Adipose tissue/adipocytes; Acute myeloid leukemia; Anaplastic; Alveolar rhabdomyosarcoma;
	Alveolar Soft Part Sarcoma; Anaplastic Thyroid Cancer;
B	B-Cell Acute lymphoblastic leukemia; Breast cancer; Basal cell carcinoma; B-cell lymphoma; Benign tissue;
	Benign glands; Benign colon tissue; Bladder cancer including melonoma; Benign rectal tissue;
	Benign essential blepharospasm; Bone marrow; BRCA1 and BRCA2; Brain Tumor;
	Breast invasive lobular carcinoma; Benign multi-cystic peritoneal mesothelioma; Breast invasive ductal carcinoma;
C	Cancer; carcinoma; Cancer staging; Carcinoma in situ; Carotid body tumor; Clear cell renal cell carcinoma;
	Carcinoid tumor; Carcinoma In Situ; Chronic granulocytic leukemia; Cervical intraepithelial neoplasia;
	Chronic inflammatory bowel disease; Chromophobe renal cell carcinoma; Cirrhosis; Cell-mediated immunity;
	Chronic myeloid leukemia; Chronic myelomonocytic leukemia; Cytomegalovirus; Comed—comedocarcinoma;
	Colectomy; Colitis; Colon polyp; Colonoscopy; Cancer-associated stroma; Coloncancer adenocarcinoma debris;
	Core biopsy; choroid plexus carcinoma; Colorectal carcinoma/cancer; Colorectal adenocarcinoma;
	Cerebrospinal fluid; Circulating tumor cell; cancerous tissue; Cutaneous T-cell lymphoma; Cerebrovascular accident;
	CC-Cervical cancer; Colitis;
D	Diffuse histolytic lymphoma; Distant recurrence; Debris or dead cell;
	Diffuse large B-cell lymphoma; Dukes staging system; Dysplasia;
E	Early-Stage invasive breast cancer; Epstein-barr virus; Esophageal cancer; Epidermal growth Factor Receptor;
	Extra skeletal myxoid chondrosarcoma; Estrogen receptor;
F	Formalin fixed paraffin embedded tissues; Fluorescence In Situ hybridization; Fibrosis;
G	Gastrointestinal stromal tumors; Gastrointestinal cancer; Gleason score; Gleason score;
H	Hand foot syndrome; Hepatocellular carcinoma; Hairy cell leukemia; Hodgkin's disease; Hyperplasia;
	Human epidermal growth factor receptor 2; Hereditary nonpolyposis colon cancer;
	Human immunodeficiency virus; Hereditary nonpolyposis colon cancer; Human T-cell Leukemia;
	Head and neck squamous cell carcinoma; Human papillomavirus; Hidradenoma papiliferum;
I	Immunohistochemistry; Invasive cancer; In Situ hybridization; Invasive ductal carcinoma of the breast;
	IInvasive carcinoma of the breast; Invasive lobular carcinoma; Inflammatory cells;
	Immune cells; Inflammatory bowel diases;
L	Langerhan's cell histiocytosis; lentigo maligna melanoma; Lung cancer; Lung squamous cell carcinoma;
	Lung adenocarcinoma; Lymph Node; Leipidic pattern adenocarcinoma; Lymphoid infiltrate;
	Leipidic predominant histologuical subtype; Lymphocytes; Liver cancer;
M	Malignant; Mouth and throat cancer; Mastectomy; Myelodysplastic syndromes; Multiple endocrine neoplasia;

	Malignant colon tissue; Malignant rectal tissue; Mucus/Mucin; Metastasis; Muscularis propria; Mucinous carcinoma;
	Micropappillary pattern; Micropappillary pattern adenocarcinoma; Micropappillary growth pattern;
	Micropappillary predominant histological subtype; Muscularis mucosa; Metastatic cells; Malignant melanoma;
	Malignant peripheral nerve Sheath tumor; Malignant rhabdoid tumor; Microsatellite instability; Microsatellite stability;
N	Normal adjacent tissue; Nevoid basal cell carcinoma Syndrome; Non-Hodgkin's lymphoma;
	Nodular melanoma; Non melanoma skin ccancer; Nasopharyngeal cancer; Node-negative breast cancer;
	Node-Positive breast Cancer; Non-Small cell lung cancer; Necrosis; Neoplasi; Neoplasm; Neutrophils;
O	Osteogenic Sarcoma; Ovarian cancer;
P	Pathologic (or Histologic) grade well differentiated; Pathologic (or Histologic) Grade moderately differentiated;
	Pathologic (or Histologic) grade poorly differentiated; pathologic (or Histologic) grade undifferentiated;
	Pathologic stage; Peripheral blood mononuclear cells; Parkinson's disease; Primary lymphoma of bone; Polyp;
	Progesterone receptor; Prostate-Specific antigen; Prostrate cancer; Prostrate cancer with gleason grade 3;
	Prostrate cancer with Gleason grade 4; Prostrate cancer with gleason grade 5; Prostrate adenocarcinoma;
	Prostatic adenocarcinoma; Papillary renal cell carcinoma; Papillary pattern adenocarcinoma;
	Pancreatic cancer; Papillary growth pattern; Papillary tumor; Papilloma;
R	Renal cell carcinoma; Renal cell carcinoma of chromophore type; Rhabdomyosarcoma;
S	Sarcoma; Squamous Cell Carcinoma; Small Cell Lung Cancer; Secondary Score;
	Synchronous cancer; Solid pattern adenocarcinoma; Solid growth pattern; Smooth muscle;
	Stromal tissue; Stromal cells; Skin cancer; Stroma associated tumor; Sialadenoma papilliferum;
T	Transitional cell carcinoma; Thrombocytopenia; classification of malignant tumors;
	Tumor; Tumor grade; Tumor-associated stroma; Tumor infiltrating lymphocytes;
	Tumor epithelial tissue; Testicular cancer;
U	Ulcerative colitis; Urinary bladder Cancer; Urinary bladder adenocarcinoma; Urinary bladder tissue;
W	White Blood cell count; Waldenstrom's macroglobulinemia;
Y	Yolk sac Tumor;