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, Pakistan, [3]The University of the Western Australia

Abstract

This paper proposes Comprehensive Pathology Language Image Pre-training (CPLIP), a new unsupervised technique designed to enhance the alignment of images and text in histopathology for tasks such as classification and segmentation. This methodology enriches vision-language models by leveraging extensive data without needing ground truth annotations. CPLIP involves constructing a pathology-specific dictionary, generating textual descriptions for images using language models, and retrieving relevant images for each text snippet via a pre-trained model. The model is then fine-tuned using a many-to-many contrastive learning method to align complex interrelated concepts across both modalities. Evaluated across multiple histopathology tasks, CPLIP shows notable improvements in zero-shot learning scenarios, outperforming existing methods in both interpretability and robustness and setting a higher benchmark for the application of vision-language models in the field. To encourage further research and replication, the code for CPLIP is available on GitHub at https://cplip.github.io/

1. Introduction

Vision Language (VL) models have substantially progressed, enhancing a broad spectrum of vision applications with their ability to understand open vocabularies and demonstrate capabilities for zero-shot transfer [9, 19, 27, 35, 36, 38]. Key to this progress is the effective alignment of visual and linguistic data, done using large datasets with paired images and text [25]. The Contrastive Language-Image Pretraining (CLIP) model exemplifies this evolution, using contrastive learning to align visual and text embeddings on a large scale [25].

Translating these advances to computational pathology, VL models have transitioned from being novel to essential, enabling the fine-tuning of datasets considerably smaller than those typically used for VL pretraining [14, 18, 21, 22]. Despite this progress, the scarcity of Whole Slide Images (WSIs) and diverse cancer morphologies poses a challenge for the zero-shot transfer capabilities of VL models, particularly for tasks like patch-based tissue recognition and WSI-level cancer subtyping, which are crucial during the inference phase [22]. Nevertheless, the successful deployment of VL models in classifying and analyzing WSIs underscores their significant role in revolutionizing the field of computational pathology.

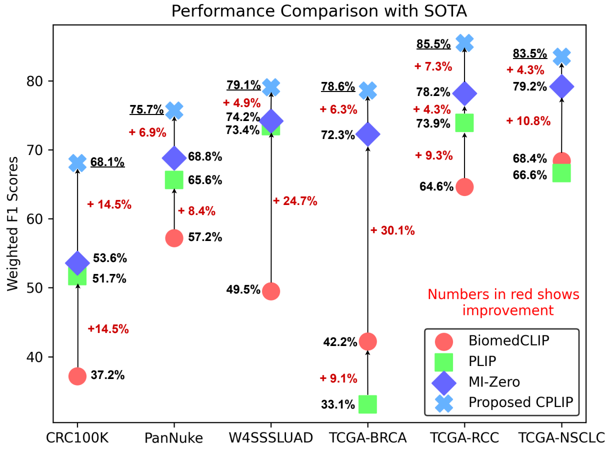

Figure 1. Comparative analysis of zero-shot classification performance between the proposed CPLIP algorithm and existing SOTA methods such as BiomedCLIP [37], PLIP [14], and MI-Zero [23]. The weighted F_1 scores demonstrate CPLIP's substantial performance enhancements across six independent histology datasets.

The use of VL models in classifying and analyzing WSIs has shown their impact on computational pathology [22]. Lu *et al.* created a dataset of 33.48K histology image-caption pairs, which helped to fine-tune the CLIP model for cancer subtyping [23]. Huang *et al.* collected about 208K histology images and texts from Medical Twitter to further fine-tune the CLIP model's ability for zero-shot classification and matching [14]. Zhang *et al.* also collected a heterogeneous dataset of 15 million image-text pairs, strengthening the CLIP model's training foundation [37].

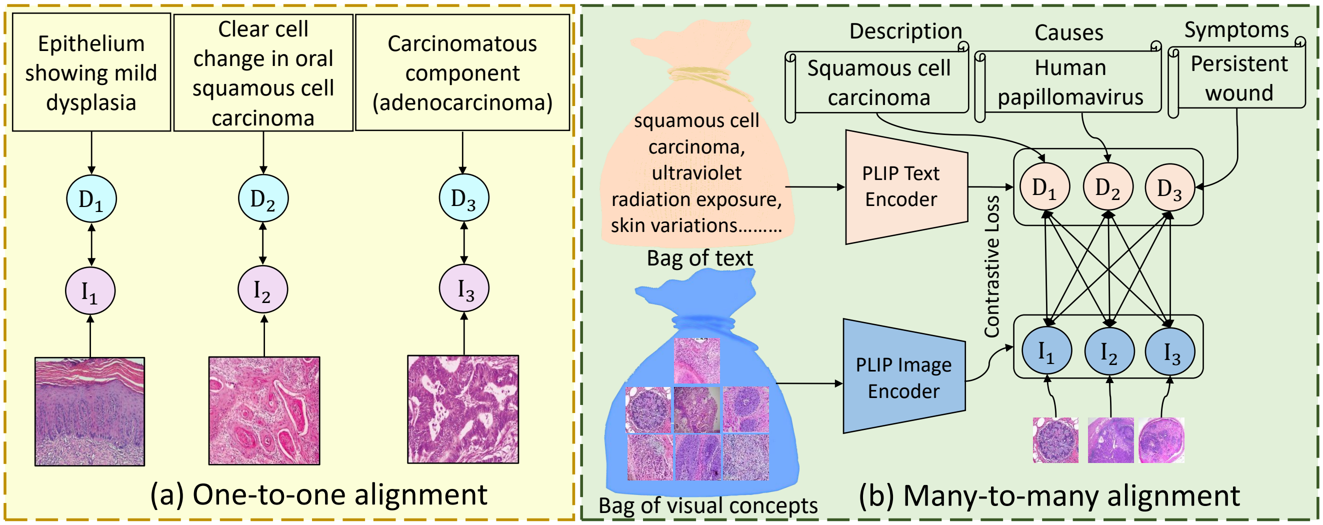
Epithelium showing mild dysplasia
Clear cell change in oral squamous cell carcinoma
Carcinomatous component (adenocarcinoma)
D_1
D_2
D_3
I_1
I_2
I_3
(a) One-to-one alignment
Description
Causes
Symptoms
Squamous cell carcinoma
Human papillomavirus
Persistent wound
squamous cell carcinoma, ultraviolet radiation exposure, skin variations.........
Bag of text
PLIP Text Encoder
Contrastive Loss
PLIP Image Encoder
Bag of visual concepts
(b) Many-to-many alignment

Figure 2. **(a)** Displays the traditional one-to-one alignment in computational pathology VL models like PLIP [14], BiomedCLIP [37], and MI-Zero [23], where each histology image is aligned with a single textual description during fine-tuning. **(b)** Our proposed approach of many-to-many alignment, where bags of correlated texts are aligned with bags of correlated histology images during fine-tuning, offers a richer, interconnected data set for model training.

Textual prompts play a crucial role in improving VL models' performance. Yet, the tendency of these models to rely on just one phrase for each histology image might limit their zero-shot classification effectiveness [14, 23, 37]. They often use simple noun-based phrases, which may ignore detailed causes and symptoms of specific cancers. Introducing richer, more detailed prompts could provide VL models with a broader range of information during training, potentially improving their ability to classify and understand various cancer types. To our knowledge, no existing histology 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 the simultaneous alignment of numerous interrelated textual and visual concepts, as depicted in Figs. 2 (a) & (b).

In this paper, 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.

To generate "comprehensive" textual prompts, we first compiled a pathology-specific dictionary cataloging various cancer types and related medical conditions, using a range of publicly available online glossaries. We then used an existing VL model [23] to select the most appropriate prompts for each histology image from this dictionary. With GPT-3 [5], we transformed the selected prompts into five unique variations and identified three main causes and symptoms for each condition. Using the Pathology Language Image Pre-training (PLIP) model [14], we matched these enhanced prompts with corresponding histology images from a Twitter dataset to enrich our visual database. The number of textual descriptions and images was capped at 17 and 21 to manage computational demands, though this limit can be adjusted according to resource availability.

Using our extensive collection of textual prompts and visual content, we generated collections—or 'bags'—of textual descriptions and images through an unsupervised and automated process. Images that match the prompts from our pathology dictionary are labeled as positive examples, while mismatches are negative. These collections are then used to fine-tune the CLIP model by adjusting the model's embeddings to align similar (positive) concepts and push away dissimilar (negative) ones. This method is aimed to enhance class-agnostic representations (refer to Fig. 1). Our resulting fine-tuned model, called Comprehensive PLIP (CPLIP), is suited for various downstream zero-shot classification tasks.

This approach aligns with trends in AI that enhance interaction between language and visuals, much like VISPROG [12], which translates language instructions into visual task actions. Similarly, our proposed CPLIP model integrates detailed textual and visual information to improve understanding in computational pathology. In summary, our contributions include:

- Compilation of a dedicated dictionary for pathology-related prompts to facilitate the organized collection and application of comprehensive textual descriptions, improving model training and evaluation (Sec. 3.1).

- Development of comprehensive textual descriptions paired with multiple visual concepts to better align text and image embeddings (Secs. 3.2 & 3.3).
- Advocacy for collective alignment of multiple textual descriptions and visual concepts (Sec. 3.5).
- Demonstrated superior zero-shot performance by our model on different datasets, highlighting the benefits of the "comprehensiveness" approach to boosting VL models for classification and segmentation in computational pathology (Sec. 4).

2. Related Work

In the effort to advance computational pathology through various tasks like histology image classification, segmentation, and survival prediction, numerous methods have been proposed [7, 29]. These methods can be broadly categorised as weakly-supervised [15, 26], self-supervised [17, 32], and Vision-Language (VL) supervised [14, 23, 37].

(i) Weakly-supervised Learning Methods (WSL) use data with labels at a broad level, without needing detailed annotations for every instance. In computational pathology, Multiple Instance Learning (MIL) has evolved as a popular paradigm for WSI classification. Examples include ABMIL [15], TransMIL [26], DSMIL [26], CLAM [20], and DTFD-MIL [34]. *In contrast to this paradigm, our VL-based algorithm does not require any label during the training rather it uses pathology-specific language supervision.*

(ii) Self-supervised Learning Methods in computational pathology learn from the data itself without using labels, using pretext tasks to boost downstream task performance. Key in this area is contrastive learning-based methods, which focus on distinguishing between similar and contrasting instances within the data. By using contrastive loss, these methods train models to discern augmentation-invariant features crucial for tasks like classification and anomaly detection. Examples include CTransPath for histology image classification [32], H2T [31], HIPT [6], and [17]. These techniques help models capture essential inherent data characteristics, enhancing performance on various computational pathology applications. *Our approach goes beyond contrastive learning methods by not only aligning bags of images but also aligning bags of texts and additional strategies to enhance model performance.*

(iii) Learning with Pathology Language Supervision Methods integrate textual descriptions with visual data to pre-train deep models. Adhering to the conventional VL model training approach, these methodologies leverage paired visual-textual data within a contrastive learning framework to ensure that representations of similar visual-textual concepts are drawn closer together, while divergent ones are distanced [7, 22, 25]. Recently, the VL paradigm has been extended to zero-shot classification and segmentation tasks, introducing models like PLIP [14], CONCH [22], MI-Zero [23], and BiomedCLIP [37]. These innovations have brought forth new datasets containing descriptions of histology images and languages to pre-train architectures resembling CLIP [25]. A limitation of these models is their potential inability to generalize well across different datasets due to the training dataset-specific biases consisting of paired textual-visual concepts. Also most of these approaches primarily focus on aligning single textual and visual representations. In contrast, we propose the engagement of comprehensive visual and textual data to concurrently align multiple correlated positive visual-textual concepts. We argue that such an expansive and robust alignment significantly elevates performance across a spectrum of computational pathology tasks.

3. Proposed Methodology

In this work, we propose the Comprehensive Pathology Language Image Pre-training (CPLIP) algorithm. This algorithm effectively uses a collection of unlabeled histology images, paired with a predefined comprehensive pathology prompt dictionary, to fine-tune the CLIP model without any ground truth annotations (neither at the image level nor at the text level). The purpose is to tailor CLIP to a diverse range of histology data gathered from various sources. This enhances its ability for zero-shot transfer across different computational pathology tasks, especially for unfamiliar tissue categories not encountered during the training phase. We represent the comprehensive pathology prompts dictionary as V and denote the collection of unlabeled histology images as $H = \{h_j\}_{j=1}^{n_h}$, where n_h indicates the total count of these histology images.

Fig. 3 illustrates the process of constructing the bag of textual descriptions and the bag of visual concepts within our CPLIP framework. It depicts the primpary phases, including the construction of a predefined pathology prompt dictionary and the aggregation of corresponding textual descriptions, followed by the formation of visual concepts. These elements are integral to our many-to-many contrastive learning approach, which seeks to align positive visual-textual pairs and separate negative ones. The details of these processes are discussed in the following sections.

3.1. Predefined Pathology Dictionary (Fig. 3 A(a))

The ARCH dataset is the only publicly accessible histology image-caption pairing [11]. This dataset has been used in MI-Zero [23]. This method, however, restricts them to paired image-text data, which might not be comprehensively available to the public. To address this limitation, we propose a set pathology prompt dictionary. This serves as a foundational prompt to extract more comprehensive images and textual descriptions in subsequent phases.

We have created a strong dictionary tailored to histology

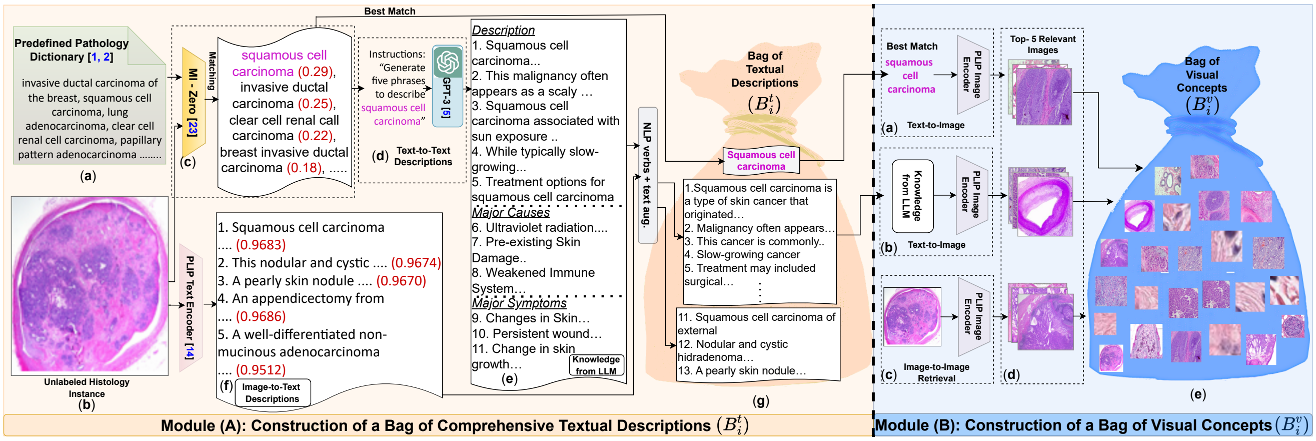
Predefined Pathology Dictionary [1, 2]
invasive ductal carcinoma of the breast, squamous cell carcinoma, lung adenocarcinoma, clear cell renal cell carcinoma, papillary pattern adenocarcinoma ....
(a)
Unlabeled Histology Instance
(b)
MI - Zero [23]
Matching
(c)
Best Match
squamous cell carcinoma (0.29), invasive ductal carcinoma (0.25), clear cell renal call carcinoma (0.22), breast invasive ductal carcinoma (0.18), ...
Instructions: "Generate five phrases to describe squamous cell carcinoma"
GPT-3 [5]
(d) Text-to-Text Descriptions
PLIP Text Encoder [14]
1. Squamous cell carcinoma .... (0.9683)
2. This nodular and cystic .... (0.9674)
3. A pearly skin nodule .... (0.9670)
4. An appendicectomy from .... (0.9686)
5. A well-differentiated non-mucinous adenocarcinoma .... (0.9512)
(f) Image-to-Text Descriptions
Description
1. Squamous cell carcinoma...
2. This malignancy often appears as a scaly …
3. Squamous cell carcinoma associated with sun exposure ..
4. While typically slow-growing...
5. Treatment options for squamous cell carcinoma
Major Causes
6. Ultraviolet radiation....
7. Pre-existing Skin Damage..
8. Weakened Immune System…
Major Symptoms
9. Changes in Skin…
10. Persistent wound…
11. Change in skin growth…
Knowledge from LLM
(e)
NLP verbs + text aug.
Bag of Textual Descriptions
(B_i^t)
Squamous cell carcinoma
1.Squamous cell carcinoma is a type of skin cancer that originated…
2. Malignancy often appears…
3. This cancer is commonly..
4. Slow-growing cancer
5. Treatment may included surgical…
11. Squamous cell carcinoma of external
12. Nodular and cystic hidradenoma…
13. A pearly skin nodule…
(g)
Module (A): Construction of a Bag of Comprehensive Textual Descriptions (B_i^t)
Best Match
squamous cell carcinoma
PLIP Image Encoder
(a) Text-to-Image
Knowledge from LLM
PLIP Image Encoder
(b) Text-to-Image
PLIP Image Encoder
(c) Image-to-Image Retrieval
Top- 5 Relevant Images
(d)
Bag of Visual Concepts
(B_i^v)
(e)
Module (B): Construction of a Bag of Visual Concepts (B_i^v)

Figure 3. Diagram outlining the construction of comprehensive textual descriptions and visual concept bags. (A) illustrates the construction process of the textual description bag, while (B) shows the procedure for constructing the visual concept bag. Within (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. Within (B), there are also three primary steps: (a) using PLIP to identify the best-matching images, (b) leveraging PLIP to enrich the histology images of the best-matched textual descriptions, and (c) employing the PLIP to retrieve relevant histology images of the input unlabeled histology image.

descriptions, which includes terms commonly used by expert pathologists to describe various cancer forms, related medical conditions, and their prognoses through histology images. To generate this resource, we merged cancer glossaries from esteemed institutes [1, 2] manually refining the collected data to form a more precise pathology-specific dictionary. Our experiments compare the effectiveness of these two vocabularies, assessing the outcomes of each. The combined dictionary holds 500 varied prompts, incorporating 1,500 terms covering the range of cancer types and morphologies for diagnosis. After refinement (cleaning and filtering), the dictionary has 200 wide-ranging and in-depth prompts totaling 700 terms. This refinement process first removes irrelevant prompts, like those not directly connected to a histology image. It then omits non-histopathology prompts, sidestepping those related to radiology, X-rays, CTs, and so on. Every prompt is thereafter denoted with a suitable acronym and corresponding description. The refined predefined prompts dictionary is designed to cover major cancer types and morphologies across various tissue types. We have provided it as supplementary material in this paper and intend to release it to the public.

3.2. Building a Comprehensive Textual Descriptions Bag (Fig. 3 A(g))

Given the collection of input unlabelled histology images, denoted as H, and the predefined pathology prompts dictionary, V, we generate a detailed textual description bag, B_i^t, for each image $h_j \in H$. This process uses three distinct textual sources: MI-Zero [23], GPT-3 [5], and PLIP [14]. Each source is elaborated on below.

3.2.1 Matching with MI-Zero (Fig. 3 A(c))

While the widely used visual text encoder CLIP is trained on generic data, our work necessitates a domain-specific VL encoder. With limited options available, we opted for the MI-Zero model [23]. This model, recently launched, is trained on matched histopathological image-caption data. MI-Zero includes a visual encoder, represented as $f(\cdot;\theta)$, and a text encoder, $g(\cdot;\phi)$, both of which compute image and text embeddings, respectively. For a given histology image h_j, we identify its most related prompts from the predefined dictionary V using the formula:

$$\hat{v}_i = \underset{v_i \in V}{\operatorname{argmax}} \, \text{sim}(f(h_j), g(v_i)), \tag{1}$$

Here, $\text{sim}(x, t) = x^\top t/(||x|| \; ||t||)$ denotes the cosine similarity measure. The resulting $\hat{v}_i$ is then added to the textual descriptions bag, B_i^t. Fig. 3 (A) provides a visual representation of this process. It begins with the predefined pathology prompt dictionary (shown in Fig. 3 A(a)) and an unlabelled histology image (Fig. 3 A(b)). From here, we identify the top five matching prompts (Fig. 3 A(c)). Of these, only the best-matching text, termed "squamous cell carcinoma", is chosen and added to the bag B_i^t (See Fig. 3 A(g)). For more examples of closely matched prompts, refer to our supplementary material.

3.2.2 GPT-3 for Comprehensive Textual Descriptions (Fig. 3 A(d))

To derive multiple descriptions of the top-ranked textual prompt from the previous process (Sec. 3.2.1), we can turn to Large Language Models (LLM) like GPT-3 [5]. Such models have demonstrated strong capabilities in various linguistic tasks [33]. By inputting the highest ranked prompt

$\hat{v}_i$ into LLM, we ask it to produce five alternate descriptions based on its extensive linguistic knowledge. Additionally, we generate three primary etiologies/causes and three dominant symptoms related to the top-ranked prompt using the LLM. An example presented in Fig. 3 A(e) reveals five unique descriptions for the top-rated prompt “squamous cell carcinoma”. Some descriptions include “squamous cell carcinoma is a common form of skin cancer” and “squamous cell carcinoma malignancy often appears as a scaly, red patch...” and so on. The model also provides potential causes like “prolonged exposure to ultraviolet radiation” and noticeable symptoms like “skin alterations” and “lingering wound”. Together with the identified pathology prompt “squamous cell carcinoma”, this brings forth twelve varied textual descriptions that will be considered during the formulation of our bag B_i^t. Moreover, our text augmentation includes the lemmas of verbs, helping the model to treat different verb forms as the same action. Additional illustrations are made available in the supplementary material.

3.2.3 Image-to-Text Description with PLIP (Fig.3 A(f))

In this phase, we use the PLIP model to identify relevant text descriptions linked to given unlabeled histology images, pulling information from the vast Medical Twitter dataset [14]. From this, we select the top five most appropriate descriptions and add them to our textual bag, B_i^t. The PLIP model consists of both visual and text encoders, specifically adapted based on the large-scale medical Twitter dataset. With PLIP’s assistance, we can integrate descriptions from a variety of sources into our textual bag, B_i^t. Fig. 3 A(f) displays the top five descriptions matched to the unlabeled histology image shown in Fig. 3 A(a).

3.3. Compiling Visual Concepts Bag (Fig. 3 B(e))

The visual concepts repository, denoted as B_i^v, is constructed based on the comprehensive textual descriptions sourced from the textual bag B_i^t and the unlabeled histology image h_j. This process consists of two primary stages, as depicted in Fig. 3 (B).

3.3.1 Image Retrieval with PLIP Based on Textual Prompts (Fig. 3 B(a)-(b))

Starting with a top-matched prompt from MI-Zero matching, we identify several histology images that match this prompt using PLIP’s image and text encoders. Similarly, using a set of textual descriptions from LLM, we find related histology images that go with each description through the PLIP model. It is important to note that our approach uses only the pre-trained PLIP model without any extra fine-tuning. For example, with the prompt “squamous cell carcinoma” as our top match, and using textual information from LLM (as shown in Fig. 3 A(e)), we were able to identify a total of 16 images that were relevant.

3.3.2 PLIP Image-to-Image Retrieval (Fig. 3 B(c))

With the unlabelled histology image, h_j, we retrieve the five most related images using PLIP’s image encoder from the Medical Twitter dataset. When these are added to B_i^v, the total comes to 21 visual concepts. The top five images, as an example (Fig. 3 B(c)), can be viewed in the supplementary material.

3.4. Textual and Visual Bags Pruning

Considering the textual descriptions in B_i^t come from various sources, there is a chance some may not be as relevant. To improve the B_i^t bag quality, we make sure each description $t_{i,n} \in B_i^t$ closely matches with input image h_j, exceeding a specific similarity value $\text{sim}(f(h_j), g(v_i)) \geq \delta_t$. Adjusting this δ_t value can either reduce the number of descriptions in B_i^t (if the value is higher) or keep most of them (if it is lower). Since B_i^v is constructed using the pruned textual bag and the PLIP model, the pruning applied on the B_i^t consequently reflects in the B_i^v. Please note no further pruning is applied on bag B_i^v.

3.5. MIL-based Contrastive Loss (Fig. 2 (b))

To fine-tune the PLIP model, we use the Multiple Instance Learning-Noise Contrastive Estimation (MIL-NCE) loss introduced in [24]. Contrary to the original MIL-NCE design that aligns a single positive text with a single positive video, our algortihm CPLIP connects a bag of text, B_i^t, with a corresponding set of visual bags, B_i^v (the specific sequence of items from the bags is inconsequential). This approach facilitates the association of multiple textual descriptions with multiple histology images. Our defined MIL-NCE loss function is presented as follows:

$$\mathcal{L} = -\frac{1}{B}\sum_i log\left[\frac{\sum_m \sum_n \exp\left(f(v_{i,m})^\top g(t_{i,n})/\sigma\right)}{\sum_m \sum_j \sum_n \exp\left(f(v_{i,m})^\top g(t_{j,n})/\sigma\right)}\right], \tag{2}$$

Here, $v_{i,m} \in B_i^v$, $t_{j,n} \in B_j^t$, $0 < n \leq n_{bag}$, and $0 < m \leq m_{bag}$. m_{bag} represents the size of B_i^v, while n_{bag} denotes the size of B_i^t. B and σ indicate the batch size and the constant temperature parameter, respectively.

3.6. Zero-shot Transfer for Histology Landscape

Radford *et al.* have introduced a method that uses prompts for zero-shot classification [25]. In this method, class names are converted into prompts by attaching them to specific keyword templates. For instance, the class name “Tumor Adenocarcinoma” is expanded using the template “An H & E image of {}”. Subsequently, the trained text encoder calculates the embeddings of these prompts. Meanwhile, the trained visual encoder deduces the embeddings of test images. These embeddings are normalized using ℓ_2, and their

Table 1. Ablations 1-3: Zero-shot classification performance comparison in terms of weighted F_1 score using different heterogeneous textual descriptions. 95% Confidence Interval (CI) is included in parentheses.

Ablation Study	D500+GPT-3+PLIP	D200+GPT-3+PLIP	D200 only	D200+GPT-3	GT+GPT-3+PLIP
CRC100K	0.804(0.791,0.815)	0.844(0.833, 0.856)	0.697(0.682,0.704)	0.774(0.752,0.703)	**0.861(0.852,0.874)**
DigestPath	0.842(0.833, 0.859)	0.903(0.891, 0.915)	0.734(0.707,0.764)	0.831(0.820,0.834)	**0.912 (0.904,0.922)**
SICAP	0.441(0.401,0.485)	0.511(0.498, 0.526)	0.292(0.276,0.317)	0.422(0.375,0.475)	**0.533(0.508,0.571)**
W4SSSLUAD	0.801(0.772,0.835)	0.882(0.876, 0.894)	0.644(0.605,0.683)	0.716(0.685,0.743)	**0.891(0.876,0.916)**
PanNuke	0.761(0.744,0.786)	0.811(0.799, 0.827)	0.685(0.613,0.749)	0.761(0.753,0.772)	**0.841(0.815,0.873)**

Table 2. Ablation 5: Zero-shot classification performance comparison in terms of weighted F_1 score for a bag of text vs. a bag of visual concepts. 95% CI is included in parentheses.

Ablation Study	Text bag (B^t)	Visual bag (B^v)	Proposed
CRC100K	0.761(0.753,0.774)	0.744(0.723,0.765)	**0.844(0.833,0.856)**
DigestPath	0.854(0.831, 0.872)	0.861(0.852,0.871)	**0.903(0.891,0.915)**
SICAP	0.477(0.465,0.487)	0.471(0.451,0.495)	**0.511(0.498,0.526)**
W4SSSLUAD	0.772(0.752,0.793)	0.786(0.772,0.796)	**0.882(0.876,0.894)**
PanNuke	0.766(0.734,0.795)	0.756(0.723,0.785)	**0.811(0.799,0.827)**

similarity is measured using the cosine similarity measure. The labels of the test images are determined based on the highest similarity scores. Given the variance in the performance of different prompts, we expand the prompt generation process. We use a set of templates tailored for pathology and introduce alternative names for each class, drawing inspiration from earlier studies [22, 23]. When making inferences, the various prompts for each class are combined by averaging their embeddings. Our experiments present results both with and without the merging of prompts.

4. Experiments

We conduct several experiments to evaluate the proposed CPLIP algorithm, including tile-level zero-shot classification, WSI-level zero-shot classification, and zero-shot segmentation of gigapixel WISs. For the tile-level zero-shot classification, we use five independent datasets: CRC100K [16], WSSS4LUAD [13], PanNuke [11], DigestPath [8], and SCIAP [28]. For the WSI-level zero-shot classification, we use four datasets: CAMELYON-16 (CAM16) [4] and others from The Cancer Genome Atlas (TCGA) including BRCA, RCC, and NSCLC [30]. Finally, for the zero-shot segmentation, we use the SICAP and DigestPath datasets. Through these diverse experiments spanning tiles, WSIs, and segmentation tasks, we comprehensively assess the performance of the proposed CPLIP method.

4.1. Training and Implementation Details

In histopathology, the ARCH dataset [10] is the only widely available image-text paired dataset, containing 8,617 pairs from clinical and research pathology articles. We fine-tuned our CPLIP algorithm on this dataset, extending it to around 180,000 images and 146,000 textual descriptions without using their paired texts. This unpaired many-to-many image-text alignment is a novelty compared to the paired data approach used by MI-Zero [23]. Our fine-tuning process involved various architectures, leveraging domain-specific and general models, with modifications to suit our many-to-many alignment needs. We used a batch size of 256 for 50 epochs, applying specific filtering thresholds to refine the data further. While single prompts were used for reporting results, additional details on the use of merged prompts and further implementation details are provided in the supplementary material section.

4.2. Datasets and Evaluation Metrics

We used nine independent publicly available **computational pathology datasets** for classification and segmentation tasks (more detailed descriptions of each dataset are provided in the supplementary material), spanning diverse cancer types and image modalities including **(i)** CRC100K [16] colorectal cancer dataset used for zero-shot tile classification on 7,180 test images across nine tissue types; **(ii)** WSSS4LUAD [13] lung adenocarcinoma dataset used for zero-shot tumor vs. normal classification on 3,028 test images; **(iii)** SICAP [28] prostate cancer dataset used for zero-shot classification on 2,122 test images with 4 Gleason pattern labels; **(iv)** PanNuke [11] diverse tissue dataset used for zero-shot tumor vs. normal classification on 1,888 test images with 19 tissue types; **(v)** DigestPath [8] colonoscopy tissue dataset used for zero-shot tumor vs. normal tile classification on 18,814 test images; **(vi)** Camelyon 16 (CAM16) [4] breast cancer dataset used for zero-shot slide classification on 130 test slides; and **(vii-ix)** TCGA [30] invasive BRCA, RCC, and NSCLC datasets used for zero-shot slide classification on 75 slides per class. In summary, these diverse ranges of computational pathology datasets are used to evaluate zero-shot classification and segmentation performance across testing sets ranging from thousands of image tiles to hundreds of WSIs. Our **evaluation metrics** include balanced accuracy, weighted F_1 score, and AUCROC for classification tasks, and the Dice score, precision, and recall for segmentation tasks, in line with current SOTA VL methods [14, 22, 23]. Balanced accuracy is calculated by averaging the recall of each class.

4.3. SOTA Methods for Comparison

We compared the performance of our proposed CPLIP algorithm with several recently proposed SOTA methods on zero-shot classification and segmentation tasks for histopathology images. We included five recently proposed

Table 3. Ablation 4: Zero-shot classification performance in terms of bags pruning using weighted F_1 score with 95% CI.

Matching	ratio δ_t	CRC100K	DigestPath	SICAP	W4SSSLUAD	PanNuke
MI-Zero matching	100%	0.806(0.791,0.813)	0.871(0.867,0.884)	0.446(0.402,0.485)	0.871(0.864,0.885)	0.798(0.775,0.817)
MI-Zero matching	90%	**0.844(0.833,0.856)**	**0.903(0.891,0.915)**	0.488(0.474,0.493)	**0.882(0.876,0.894**)	**0.811(0.799,0.827)**
MI-Zero matching	70%	0.833(0.821, 0.841)	0.896(0.861, 0.928)	**0.511(0.498,0.526)**	0.880(0.861, 0.890)	0.804(0.791,0.813)
MI-Zero matching	50%	0.829(0.814,0.838)	0.883(0.864,0.905)	0.507(0.472, 0.534)	0.875(0.866,0.886)	0.805(0.775, 0.836)
MI-Zero matching	30%	0.827(0.831,0.858)	0.881(0.876,0.898)	0.501(0.485,0.525)	0.873(0.854,0.895)	0.803(0.786,0.825)

VL-based methods in our comparison: baseline CLIP [25], PLIP [14], MI-Zero [23], BiomedCLIP [37], and CONCH [22]. To ensure a fair comparison, we used the official source code for all methods and kept the same settings for testing splits and inference prompts, except for CONCH, whose source code is not yet available.

4.4. Ablation Studies

All ablation studies use CTransPath [32] as the image encoder and BioClinicalBert [3] to initialize the text encoder, with performance reported using merged prompts. For more details and ablation studies, see supplementary material.

1. Cleaned vs. Uncleaned Pathology Prompts dictionary. This experiment compares the performance of zero-shot classification using the original unsupervised pathology prompts dictionary consisting of 500 prompts (D500) vs. a manually cleaned pathology prompts dictionary containing 200 prompts (D200) (see Sec. 3.1). As shown in Table 1, D200+GPT-3+PLIP achieved better performance on five datasets compared to D500+GPT-3+PLIP. *This indicates that a smaller, curated dictionary of 200 cleaned pathology prompts yields better zero-shot classification results than a larger, uncleaned noisy set of 500 prompts.*

2. Effect of paired image-text supervision. This experiment removed the pathology prompts dictionary step and used the ARCH paired text as the best match prompt in Sec. 3.1. The paired text data was then used to construct the textual bag using GPT-3 and PLIP text encoder to obtain a similar textual bag as in Sec. 3.3. The zero-shot classification performance of this strategy (GT+GPT-3+PLIP) is also shown in Table 1. The results showed that ground truth text-based results were better than the unsupervised dictionary results. *This indicates that using ARCH's paired text data to construct textual bags via GPT-3 and PLIP text encoder, as opposed to an unsupervised dictionary, improves zero-shot classification performance.*

3. Importance of heterogeneous textual and visual resources. We conducted experiments using only D200 pathology dictionary (using a single best match prompt and a single image), D200+GPT-3 (using 12 textual descriptions and 12 images), and D200+GPT-3+PLIP (using 17 textual descriptions and 21 images). *The results in Table 1 show that adding more textual resources during training improves performance on all datasets.*

4. Effect of Bags Pruning. In this experiment, the textual bags (B^t) were pruned to retain 90%, 70%, 50%, and 30% of the best matching textual descriptions with the input image using cosine similarity (see Sec. 3.4). The corresponding visual bags (B^v) were also pruned subsequently. A 100% bag means no pruning, and it may contain some noisy text. As shown in Table 3, the best zero-shot classification performance over four datasets was observed for $\delta_t = 90\%$. *Further pruning reduced performance due to data reduction, which resulted in reduced heterogeneity.*

5. Which Bag is more important? We conducted two experiments to compare the importance of the textual and visual bags for contrastive training. In the first experiment, we used a bag of text along with the input image ($B^t_j + h_j$). In the second experiment, we used only the bag of visual concepts B^v. Both experiments observed performance degradation compared to the proposed $B^t + B^v$ based training, as shown in Table 2. *This suggests that both the textual and visual bags are important for achieving good performance.*

4.5. Tile-Level Zero-shot Classification Results

We conducted tile-level zero-shot classification on five distinct datasets, evaluating only their test splits. The outcomes, detailed in Table 4, benchmark our CPLIP algorithm against current SOTA VL-based methods across balanced accuracy, weighted F_1, and AUROC scores, all based on a single prompt. Comprehensive results using merged prompts are available in the supplementary material. Our CPLIP model consistently outperformed others in both single and merged prompt scenarios on all datasets. The CONCH algorithm was the next best, showing strong results on the CRC100K and SICAP datasets, though its performance on the other datasets was not documented.

CPLIP notably enhanced performance compared to CONCH, with gains of 13.5% in balanced accuracy, 13.9% in weighted F_1, and 2.1% in AUROC for the CRC100K dataset using single prompts. For the SICAP dataset, the improvements were 1.7% in balanced accuracy and 14.30% in weighted F_1. Against MI-Zero/PLIP on the DigestPath and PanNuke datasets, CPLIP's enhancements were (1.3%, 4.5%, 2.2%) and (2.2%, 6.90%, 3.0%), respectively. On WSSS4LUAD, CPLIP outperformed MI-Zero by 5.6% in balanced accuracy, 4.9% in weighted F_1, and 3.1% in AUROC. *These significant performance improvements over SOTA methods demonstrate the advantages of our CPLIP algorithm.*

Table 4. Tile-level zero-shot classification performance comparison using single prompt in terms of balanced accuracy, weighted F_1, and AUROC scores with other SOTA methods across five datasets. The CPLIP algorithm outperforms existing models. For the WSSS4LUAD dataset, CONCH used a different split, denoted by an asterisk (*).

Methods	CRC100K	DigestPath	SICAP	WSSS4LUAD	PanNuke
CLIP baseline [25]	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 [37]	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 [14]	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 [23]	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 [22]	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**

Table 5. WSI-level zero-shot classification performance comparison using single prompts, in terms of balanced accuracy, weighted F_1, and AUROC on four datasets. Both our Out-of-Domain (OoD) CPLIP$_1$ and In-Domain (InD) CPLIP$_2$ outperform across all metrics.

Models (Single prompts)	Image encoder pretraining	Text encoder pretraining	CAM16	TCGA-BRCA	TCGA-RCC	TCGA-NSCLC
CLIP baseline [25]	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 [37]	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 [14]	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 [23]	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 [22]	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**

4.6. WSI-Level Zero-shot Classification Results

For zero-shot classification of gigapixel WSIs, we adopted an approach akin to MI-Zero [23]. We binarized each WSI to distinguish tissue from the background using the OTSU method and extracted N number of tiles each with 224×224 pixels. Each tile's embedding was obtained via the CPLIP image encoder and ℓ_2-normalization. We then calculated cosine similarities between tile embeddings and text embeddings, producing C similarity scores per tile. These were aggregated using top-K pooling, averaging the highest K scores per class to determine the slide-level class prediction, with K chosen from 1, 5, 10, 50, 100 based on best performance metrics (i.e., the highest balanced accuracy, weighted F_1, and AUROC scores for classification tasks).

Table 5 compares our CPLIP algorithm's zero-shot performance with SOTA VL models on CAM16, TCGA-BRCA, TCGA-RCC, and TCGA-NSCLC datasets, using a single prompt. Detailed results with merged prompts are in the supplementary material. CPLIP's performance was also assessed with various out-of-domain and in-domain encoders. CPLIP consistently outperformed in-domain VL models like PLIP, BiomedCLIP, MI-Zero, and CONCH. For instance, CPLIP$_2$ in-domain zero-shot balanced accuracy reached 59.10% for lymph node metastasis in CAM16, surpassing MI-Zero by 13.50%. In NSCLC and RCC subtyping, CPLIP$_2$ achieved balanced accuracies of 85.40% and 84.40%, respectively, outperforming CONCH and MI-Zero by margins up to 5.20%. Notably, in the BRCA subtyping task, CPLIP$_2$ achieved an 82.40% balanced accuracy, significantly ahead of CONCH and MI-Zero by 18.10% and 4.30%, respectively. *These results highlight CPLIP$_2$ SOTA performance in cancer subtyping using zero-shot learning.*

4.7. Zero-shot Segmentation of Gigapixel Images

We also performed zero-shot slide-level segmentation similar to CONCH [22] using the SICAP (31 WSIs) and DigestPath (250 large images) datasets. Overall, CPLIP outperformed other VL methods in both datasets by a significant margin demonstrating the advantages of heterogeneous textual descriptions and histology images. For further details, consult our supplementary material.

5. Conclusion

Existing visual learning (VL) models in computational pathology require paired image and text data for zero-shot learning. In contrast, we propose an algorithm that enables unpaired alignment of image and textual data for zero-shot learning in histopathology. We construct a comprehensive bag of textual descriptions using heterogeneous sources including cancer glossaries, GPT-3, and off-the-shelf VL models. These are used to build a corresponding bag of visual concepts. A bag-based contrastive learning approach then aligns the textual and visual concepts semantically. Extensive experiments on nine independent datasets demonstrate the superior zero-shot classification and segmentation performance of our proposed Comprehensive Pathology Language Image Pre-training (CPLIP) algorithm compared to SOTA VL models. Our framework is inherently translational to other applications and, in the future, we aim to develop a comprehensive pathologyGPT model to enhance cancer diagnosis and prognostications.

6. Acknowledgement

This publication acknowledges the support provided by the Khalifa University under Faculty Start-Up grants FSU-2022-003 Award No. 8474000401, and ASPIRE Grant: AARE20-279.

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. 7

[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. 6

[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. 2, 4

[6] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 3

[7] Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. Semi-supervised and unsupervised deep visual learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 3

[8] 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

[9] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 1

[10] 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. 6

[11] 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. 3, 6

[12] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14953–14962, 2023. 2

[13] 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

[14] 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. 1, 2, 3, 4, 5, 6, 7, 8

[15] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 3

[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. 6

[17] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 40(10):2845–2856, 2021. 3

[18] Zhengfeng Lai, Zhuoheng Li, Luca Cerny Oliveira, Joohi Chauhan, Brittany N Dugger, and Chen-Nee Chuah. Clipath: Fine-tune clip with visual feature fusion for pathology image analysis towards minimizing data collection efforts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2374–2380, 2023. 1

[19] Xin Li, Yuqing Huang, Zhenyu He, Yaowei Wang, Huchuan Lu, and Ming-Hsuan Yang. Citetracker: Correlating image and text for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9974–9983, 2023. 1

[20] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 3

[21] Ming Y Lu, Bowen Chen, and Faisal Mahmood. Harnessing medical twitter data for pathology ai. *Nature Medicine*, pages 1–2, 2023. 1

[22] 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. 1, 3, 6, 7, 8

[23] 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, 4, 6, 7, 8

[24] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 5

[25] 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. 1, 3, 5, 7, 8

[26] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. 3

[27] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983, 2023. 1

[28] 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

[29] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021. 3

[30] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015. 6

[31] Quoc Dang Vu, Kashif Rajpoot, Shan E Ahmed Raza, and Nasir Rajpoot. Handcrafted histological transformer (h2t): Unsupervised representation of whole slide images. *Medical Image Analysis*, 85:102743, 2023. 3

[32] 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. 3, 7

[33] Yuntao Wang, Yanghe Pan, Miao Yan, Zhou Su, and Tom H Luan. A survey on chatgpt: Ai-generated contents, challenges, and solutions. *arXiv preprint arXiv:2305.18339*, 2023. 4

[34] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022. 3

[35] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*, 2023. 1

[36] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 1

[37] 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. 1, 2, 3, 7, 8

[38] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1