CG-004

HistoCLIP: CLIP-Driven Multi-Label Classification for Histopathological Images

Bingyuan Bai
Japan Advanced Institute of Science and Technology (JAIST)
1-1, Asahidai, Nomi, Ishikawa, Japan
bbyuan@jaist.ac.jp

Kazunori Miyata
JAIST
1-1, Asahidai, Nomi, Ishikawa, Japan
miyata@jaist.ac.jp

***Abstract*—CLIP excels at zero-shot image–text alignment but its application to pathological multi-label classification remains underexplored. We present HistoCLIP, a novel pipeline employing a Cross-Modal Matching (CMM) module: zero-shot template-derived text embeddings serve as queries, while image features from a frozen CLIP visual encoder act as keys and values in cross-attention to refine modality alignment. On BCSS-WSSS, LUAD-HistoSeg, and PanNuke datasets, HistoCLIP achieves average accuracies of 91.21%, 95.07%, and 88.24%, outperforming other CLIP-based methods. These results demonstrate HistoCLIP's potential in digital histopathological image classification.**

***Index Terms*—CLIP, multi-label classification, pathology.**

I. Introduction

Histopathological image analysis has long been a cornerstone of digital pathology [1]–[3], where traditional convolutional neural networks (CNNs) first demonstrated remarkable success on natural image benchmarks like ImageNet. Yet, unlike single-label natural images, whole-slide tissue specimens often exhibit multiple pathological features [4], such as tumor subtypes, stromal components, and grading markers—making multi-label classification a more representative challenge in clinical practice. Today's vision transformers (ViTs) [5] and multimodal pretraining paradigms offer a powerful alternative: by dividing a histology slide into patch tokens, ViTs can learn rich spatial representations that, once fine-tuned, rival state-of-the-art CNNs. However, assembling large, fully annotated histopathology datasets for supervised pretraining remains prohibitively expensive.

The emergence of Contrastive Language–Image Pre-training (CLIP) [6] model to align image patches with free-text descriptions—provides an elegant solution. As a vision-language model, CLIP can generalize to new classes in a zero- or few-shot manner [7], which is particularly valuable when annotating histological patterns across diverse cancer types. Practically, two approaches harness CLIP for multi-label pathology classification. The first crafts and optimizes prompt templates (e.g., "a histology image showing *CLASS*") to maximize alignment scores, with optional negative prompts to reduce false positives [8]. The second augments CLIP's frozen backbone with lightweight heads—trainable modules after the visual encoder—mapping embeddings into a multi-label prediction space. Although fine-tuning classification heads often yields superior cross-domain adaptability, it incurs high annotation costs. Prompt engineering offers a way to boost zero-shot performance via prompt manipulation. This study evaluates prompt refinement strategies in cross-domain histopathology and compares them to the HistoCLIP.

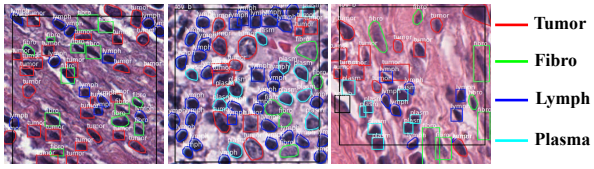

Fig. 1: Example of multi-label H&E-stained histopathology images. Three images, each exhibiting different staining intensities, contain multiple tissue and cell categories—namely, tumor cells, fibroblasts, lymphocytes, and plasma cells. Images from the NuCLS dataset [9].

We validate HistoCLIP on three large-scale multi-label histopathology datasets—BCSS-WSSS, LUAD-HistoSeg, and PanNuke—with class definitions and train–validation splits divided in an 8:2 ratio. We also refine existing zero-shot CLIP methods with diverse prompt-engineering strategies to assess their performance on histopathological images. The contributions are:

- **HistoCLIP Pipeline**: A novel cross-modal multi-label classification pipeline for histopathology.
- **Cross-Modal Matching (CMM) Module**: A module fusing text and image embeddings via cross-attention, outperforming other zero-shot approaches.
- **Extensive Comparative Evaluation**: All Experiments are conducted on two tissue-level and one cell-level dataset, directly comparing HistoCLIP against multiple CLIP-based zero-shot methods. Results show that HistoCLIP consistently outperforms these approaches, demonstrating its robustness and effectiveness.

II. Multiple Zero-Shot Strategies

CLIP's powerful zero-shot capabilities can be unlocked by framing downstream tasks as natural language prompts. Drawing inspiration from the study of prompt

for Natural Language Processing (NLP), we explored four strategies to improve zero-shot ability of Vanilla CLIP model in the multi-label classification task.

A. Baseline

For baseline method (Fig. 2), we employ the CLIP model (ViT-B/16) with OpenAI's pretrained weights for zero-shot inference, and compare it against our proposed CLIP enhancement methods: *One-vs-Rest prompting*, *Power-set prompting* and *Top-k prompting*.

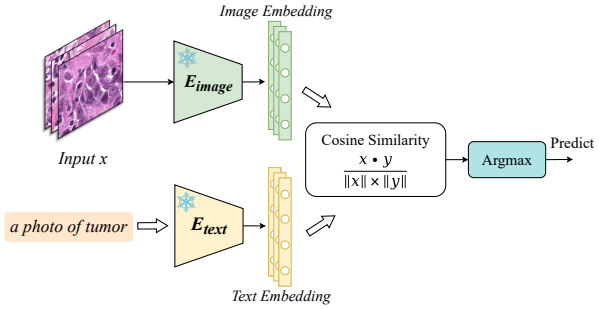

Fig. 2: The framework of baseline method. Here, x and y denote the embeddings of images and texts produced by the encoder, respectively.

The primary motivation for selecting pretrained architectures lies in their ability to encode rich, hierarchical features learned from vast image corpora. By leveraging their pretrained weights, we inherit representations honed on large-scale datasets, providing a robust initialization that accelerates convergence and improves generalization for our multi-label classification task.

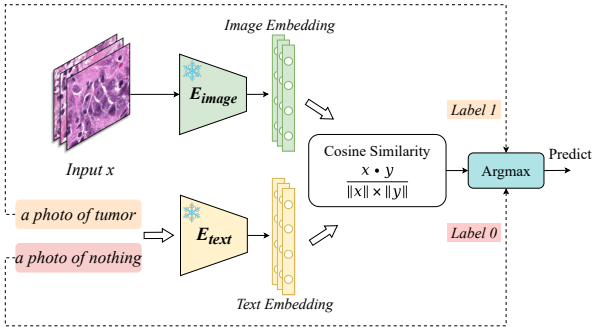

Fig. 3: The framework of One-vs-Rest prompting method.

B. One-vs-Rest prompting

For each histopathology image and each class (e.g., *tumor*, *stroma*, *lymphocyte*), we perform a binary decision by asking CLIP whether the class is present. Positive prompts take the form "a photo of a {class}", while the negative prompt is expressed as *"a photo with nothing in it."* to reduce label-specific bias. We compute the cosine similarity between the image embedding and each text embedding from CLIP's frozen encoders; if the similarity to the positive prompt exceeds that to the negative prompt, we assign label 1, otherwise 0. For example, querying *"a photo of a tumor."* on an image containing tumor tissue yields a higher similarity for the positive prompt (label=1), whereas querying *"a photo of a lymphocyte."* returns a higher similarity for the negative prompt (label=0). The framework is shown in Fig. 3.

C. Power-set prompting

In this method, for each histopathology image, we enumerate all 2^n possible subsets of n labels (e.g., *tumor*, *stroma*, *lymphocyte*) and form a single prompt by listing the classes in each subset separated by commas. We then compute the cosine similarity between the image embedding and each combined text embedding from CLIP's frozen encoders. The labels belonging to the subset with the highest similarity score are marked as 1, and all remaining labels are set to 0. See Fig. 4 for a detailed example.

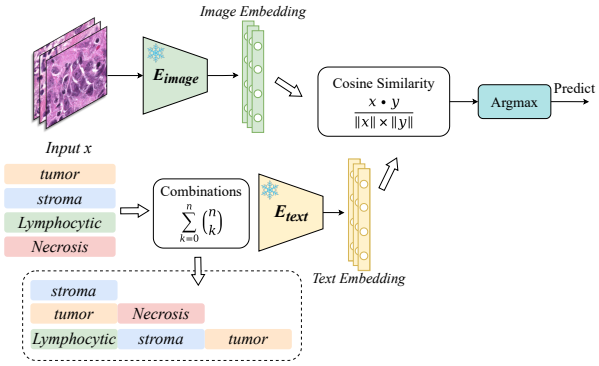

Fig. 4: The framework of Power-set prompting method.

D. Top-k prompting

Enumerating all 2^n subsets in the *Power-set prompting* scheme quickly becomes infeasible as n increases. Instead, we leverage CLIP's ranking: we first compute similarity scores for each class (e.g., *tumor*, *stroma*, *lymphocyte*) and sort them as [Top-1, Top-2, ..., Top-n]. Then, for each $k \in \{1, \ldots, n\}$, we form a prompt by listing the top-k classes (*"a photo of a Top-1, a photo of a Top-2, ..., a photo of a Top-k"*) and compare its embedding to the image. To guarantee coverage of the empty case, we include a *"a photo with nothing in it"* prompt. This reduces the search space from $O(2^n)$ to $O(n)$ while retaining CLIP's learned ranking information. Please refer to Fig. 5 for details.

E. HistoCLIP Pipeline

In contrast to the zero-shot methods, which require no additional training but offer limited gains on histopathological images due to domain mismatch, we introduce **HistoCLIP**, a three-stage pipeline. In the *feature embedding* stage, the input image is passed through CLIP's visual encoder to extract high-level representations. Next, the *Cross-Modal Matching* (CMM) stage employs a Transformer decoder to fuse class embeddings (queries)

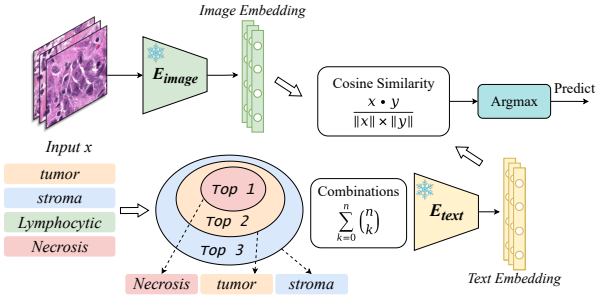

Fig. 5: The framework of Top-k prompting method.

with image features (keys and values) via multi-head attention and a feedforward network, guiding attention to the regions most relevant for each class. Finally, a lightweight post-processing step aggregates the matched embeddings to yield the final per-class scores for each sample. The data flow of pipeline is shown in Fig. 6.

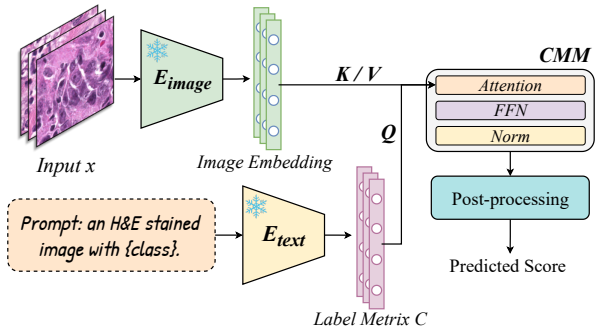

Fig. 6: The pipeline of HistoCLIP.

Given a batch of images x, we first extract visual features using CLIP's frozen visual encoder:

$$z = f_{\text{vis}}(x) \in \mathbb{R}^{B\times d}, \tag{1}$$

where B is the batch size, d the feature dimension, and $f_{\text{vis}}(\cdot)$ denotes the CLIP visual encoder producing image embeddings z. Next, we construct a zero-shot classifier matrix $\mathbf{C} \in \mathbb{R}^{d\times N}$ by encoding each of the N class labels via CLIP's text encoder and averaging over M prompt templates:

$$\mathbf{C} = \left[\bar{\mathbf{e}}^1, \ldots, \bar{\mathbf{e}}^N\right], \quad \bar{\mathbf{e}}^c = \text{Norm}\Big(\frac{1}{M}\sum_{i=1}^{M} e_i^c\Big). \tag{2}$$

Here, e_i^c is the embedding of class c under template i, M the total number of templates, and $\text{Norm}(\cdot)$ denotes L2 normalization to unit length.

These class embeddings (transposed to shape $N \times d$) serve as queries $\mathbf{Q}$, while the image features z are used as keys $\mathbf{K}$ and values $\mathbf{V}$ in a lightweight Transformer decoder. The core attention operation is

$$A = \text{softmax}\big(\mathbf{Q}\,\mathbf{K}^\top/\sqrt{d}\big)\,\mathbf{V}, \tag{3}$$

where $\sqrt{d}$ scales the dot products, and the output $A \in \mathbb{R}^{N\times B\times d}$ is further processed with residual connections and a two-layer feedforward network to yield matched embeddings T. Finally, we permute T to $\mathbb{R}^{B\times N\times d}$ and apply adaptive average pooling $\Psi(\cdot)$ along the feature dimension to obtain the predicted class scores, where $\hat{y}_{ic}$ is the logit score for sample i and class c.

$$\hat{y} = \Psi\big(\text{permute}(T)\big) \in \mathbb{R}^{B\times N}, \tag{4}$$

F. Loss and Evaluation

We train the model using the combined sigmoid and binary cross-entropy loss:

$$\mathcal{L}_{\text{BCE}} = -\big[y\log\sigma(\hat{y}) + (1-y)\log\big(1-\sigma(\hat{y})\big)\big], \tag{5}$$

where $\hat{y}$ is the raw logit, $y \in \{0,1\}$ the ground-truth label, and $\sigma(\hat{y})$ the sigmoid-activated probability.

For evaluation, we report mean accuracy (mAcc) across all C classes and N samples:

$$\begin{cases} \text{mAcc} = \dfrac{1}{C}\displaystyle\sum_{c=1}^{C}\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big(\hat{y}_{ic}^{\text{bin}} = y_{ic}\big), \\ \hat{y}_{ic}^{\text{bin}} = \mathbb{I}\big(\sigma(\hat{y}_{ic}) \geq 0.5\big). \end{cases} \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This metric averages the fraction of correct binary decisions per class.

III. Experiments

We conduct all experiments on three publicly available histopathology datasets: BCSS-WSSS [10], LUAD-HistoSeg [11], and PanNuke [12]. The first two provide 224×224 tissue-level patches labeled with four categories (Tumor, Stroma, Lymphocyte, Necrosis), while PanNuke contains 256×256 cell-level images across five nuclear types (Neoplastic, Inflammatory, Connective/Soft Tissue, Dead, Epithelial). HistoCLIP was trained on a single NVIDIA RTX 4090 GPU, system environment is Ubuntu 20.04 LTS. All three experimental settings used the same random seed and were trained for 800 epochs, including a 5-epoch warm up. We employed an initial learning rate of 1e−3, the ViT-L/14 backbone with OpenAI pretrained weights, and automatic mixed precision (AMP) for accelerated computation. The result is shown in Table I.

TABLE I: Zero-Shot mAcc Results.

Dataset	Approach	mAcc
BCSS-WSSS	Baseline	0.4266
	One-vs-Rest	0.3757
	Power-set	0.4713
	Top-K	0.4358
	HistoCLIP	**0.9121**
LUAD-HistoSeg	Baseline	0.4485
	One-vs-Rest	0.2932
	Power-set	0.4422
	Top-K	0.4738
	HistoCLIP	**0.9507**
PanNuke	Baseline	0.4107
	One-vs-Rest	0.4371
	Power-set	0.5263
	Top-K	0.4862
	HistoCLIP	**0.8824**

Across all three benchmarks, our proposed HistoCLIP demonstrates a clear advantage over both the vanilla CLIP baseline and the various zero-shots.

The baseline zero-shot inference (direct use of CLIP's pretrained ViT-L/14 without modification) yields modest mAcc scores (0.4266, 0.4485, 0.4107 on BCSS-WSSS, LUAD-HistoSeg, and PanNuke, respectively), reflecting the domain gap between natural and pathology.

The One-vs-Rest strategy—casting each class as a binary "yes/no" decision—fails to improve upon the baseline in two of three cases (BCSS-WSSS: 0.3757; LUAD: 0.2932; PanNuke: 0.4371). This suggests that simple binary prompting cannot fully overcome CLIP's bias toward natural-image contexts.

By Power-Set prompting, which exhaustively enumerates all label subsets, we see moderate gains on BCSS-WSSS (0.4713) and PanNuke (0.5263), indicating that leveraging joint label combinations can better capture co-occurrence patterns. However, its performance on LUAD-HistoSeg (0.4422) remains similar to the baseline, and the quadratic growth in prompt space limits scalability.

Top-K prompting—which ranks class probabilities and evaluates the top-k label combinations—yields mAcc scores of 0.4358, 0.4738, and 0.4862. Although this linear-complexity approach consistently outperforms both the baseline and One-vs-Rest methods, it still trails Power-Set on PanNuke and falls significantly behind HistoCLIP.

In contrast, HistoCLIP attains over 0.88 mAcc on all datasets (0.9121, 0.9507, 0.8824), more than doubling the baseline in some cases. By integrating learned cross-modal matching via a lightweight Transformer decoder, HistoCLIP effectively aligns pathological image features with class embeddings, bridging the domain gap and delivering robust, scalable multi-label classification.

IV. Conclusion and Discussion

We presented HistoCLIP, a framework that adapts CLIP to multi-label pathology classification. By extracting image features with CLIP's visual encoder, building a zero-shot classifier from text embeddings, and applying a lightweight Transformer-based Cross-Modal Matching (CMM) module, HistoCLIP significantly outperforms zero-shot prompting methods (One-vs-Rest, Power-Set, Top-K) on BCSS-WSSS, LUAD-HistoSeg, and PanNuke.

Our results highlight that simple prompting cannot fully bridge the domain gap, whereas CMM effectively aligns visual and textual embeddings to focus on pathology-relevant regions without fine-tuning the entire backbone. This design yields rapid convergence and robust mAcc improvements (up to 0.95).

Limitations include reliance on patch-level inputs and supervised CMM training. Future work will explore whole-slide aggregation, unsupervised adaptation to reduce annotation dependence, and integration of localization maps or clinical metadata for enhanced interpretability. We believe HistoCLIP's modular approach can generalize to other specialized imaging domains with scarce labels.

References

[1] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical Image Analysis*, vol. 33, pp. 170–175, 2016, 20th anniversary of the Medical Image Analysis journal (MedIA).

[2] S. Kothari, J. H. Phan, T. H. Stokes, and M. D. Wang, "Pathology imaging informatics for quantitative analysis of whole-slide images," *Journal of the American Medical Informatics Association*, vol. 20, no. 6, pp. 1099–1108, 08 2013. [Online]. Available: https://doi.org/10.1136/amiajnl-2012-001540

[3] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009.

[4] A. Echle, N. T. Rindtorff, T. J. Brinker, T. Luedde, A. T. Pearson, and J. N. Kather, "Deep learning in cancer pathology: a new generation of clinical biomarkers," *British Journal of Cancer*, vol. 124, no. 4, pp. 686–696, 2021. [Online]. Available: https://doi.org/10.1038/s41416-020-01122-x

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[7] V. S. Dorbala, G. Sigurdsson, R. Piramuthu, J. Thomason, and G. S. Sukhatme, "Clip-nav: Using clip for zero-shot vision-and-language navigation," 2022. [Online]. Available: https://arxiv.org/abs/2211.16649

[8] H. Wang, Y. Li, H. Yao, and X. Li, "Clipn for zero-shot ood detection: Teaching clip to say no," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 1802–1812.

[9] M. Amgad, L. A. Atteya, H. Hussein, K. H. Mohammed, E. Hafiz, M. A. T. Elsebaie, A. M. Alhusseiny, M. A. AlMoslemany, A. M. Elmatboly, P. A. Pappalardo, R. A. Sakr, P. Mobadersany, A. Rachid, A. M. Saad, A. M. Alkashash, I. A. Ruhban, A. Alrefai, N. M. Elgazar, A. Abdulkarim, A.-A. Farag, A. Etman, A. G. Elsaeed, Y. Alagha, Y. A. Amer, A. M. Raslan, M. K. Nadim, M. A. T. Elsebaie, A. Ayad, L. E. Hanna, A. Gadallah, M. Elkady, B. Drumheller, D. Jaye, D. Manthey, D. A. Gutman, H. Elfandy, and L. A. D. Cooper, "Nucls: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer," *GigaScience*, vol. 11, p. giac037, 05 2022.

[10] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. T. Elsebaie, L. S. Abo Elnasr, R. A. Sakr, H. S. E. Salem, A. F. Ismail, A. M. Saad, J. Ahmed, M. A. T. Elsebaie, M. Rahman, I. A. Ruhban, N. M. Elgazar, Y. Alagha, M. H. Osman, A. M. Alhusseiny, M. M. Khalaf, A.-A. F. Younes, A. Abdulkarim, D. M. Younes, A. M. Gadallah, A. M. Elkashash, S. Y. Fala, B. M. Zaki, J. Beezley, D. R. Chittajallu, D. Manthey, D. A. Gutman, and L. A. D. Cooper, "Structured crowdsourcing enables convolutional segmentation of histology images," *Bioinformatics*, vol. 35, no. 18, pp. 3461–3467, 02 2019.

[11] C. Han, J. Lin, J. Mai, Y. Wang, Q. Zhang, B. Zhao, X. Chen, X. Pan, Z. Shi, Z. Xu, S. Yao, L. Yan, H. Lin, X. Huang, C. Liang, G. Han, and Z. Liu, "Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels," *Medical Image Analysis*, vol. 80, p. 102487, 2022.

[12] J. Gamper, N. A. Koohbanani, K. Benes, S. Graham, M. Jahanifar, S. A. Khurram, A. Azam, K. Hewitt, and N. Rajpoot, "Pannuke dataset extension, insights and baselines," 2020. [Online]. Available: https://arxiv.org/abs/2003.10778