CONTRASTIVE PRETRAINING FOR COMPUTATIONAL PATHOLOGY WITH VISUAL-LANGUAGE MODELS

Qifeng Zhou, Thao M. Dang, Yuzhi Guo, Hehuan Ma, Wenliang Zhong, Saiyang Na, Jean Gao, Junzhou Huang*

Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, USA

2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI) | 979-8-3315-2052-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISBI60581.2025.10981180

ABSTRACT

In computational pathology, effectively capturing visual-language embeddings from extensive pathology image-text pairs has become increasingly crucial for diverse downstream tasks. Although prior studies have fine-tuned models like CLIP using large pathology image-text datasets, these models encounter limitations due to their separate processing of text and images, restricting their ability to capture essential cross-modal relationships critical in pathology. Recent advancements in large language models (LLMs) have led to the development of vision-language models (VLMs) that demonstrate enhanced multimodal capabilities, including stronger language comprehension and reasoning skills compared to CLIP. However, while VLMs show potential for multimodal embedding, previous efforts have primarily focused on text-based tasks, leaving their application to multimodal pathology data largely unexplored. In this work, we introduce a VLM-based framework designed to integrate and align pathology visual-language embeddings within a single model. We validate our framework's effectiveness through cross-modal retrieval on pathology image-caption datasets and zero-shot patch classification across seven pathology image datasets, demonstrating its superiority over CLIP-based models and underscoring its potential for advancing pathology research.

***Index Terms*—** computational pathology, visual-language model, contrastive learning, zero-shot learning

1. INTRODUCTION

The gold standard for diagnosing various diseases remains expert assessment by pathologists [1, 2]. Computational pathology, which uses deep learning methods, has shown great success across a range of tasks, from metastasis detection to cancer subtyping, survival prediction, and unknown primary origin site prediction [2, 3, 4]. Despite these advancements, the current paradigm typically requires large cohorts of labeled data for training specific models for individual tasks. Given the diversity of diseases and the labor-intensive nature of labeling, it is impractical to train separate models for each pathology task [3]. Nowadays, numerous pathology image-text pairs are available from resources like online databases and textbooks, which capture valuable domain-specific visual and textual information [5, 6]. Leveraging these image-text pairs to obtain multimodal embeddings is crucial, as such embeddings can support a variety of downstream tasks, potentially bypassing the limitations of fully supervised models [3].

The CLIP model [7] has demonstrated remarkable effectiveness in natural images by well-aligning visual-language embedding spaces. This success has been transferred to the field of computational pathology, where recent studies [6, 3] have fine-tuned CLIP [7] using extensive pathology image-text pairs crawled from public websites, which can capture a better pathology visual-language embeddings for several downstream zero-shot tasks. However, despite these advancements, CLIP [7] has inherent limitations [8]. Its separate processing of text and images restricts the model's ability to fully capture the intricate relationships between these modalities, which are also essential in the pathology context.

With the success of LLMs, VLMs extend LLMs to handle multimodal information [9]. Compared to CLIP [7], VLMs exhibit enhanced language comprehension, reasoning abilities, and multimodal understanding, along with improved instruction-following capacities. However, the application of VLMs to multimodal embedding representation remains largely unexplored [8]. To address these challenges, recent research [8], such as E5-V [8], has explored the potential of covering LLMs and VLMs into embedding models. Notably, E5-V [8] has proposed fine-tuning VLM with specific prompts on text data to represent multimodal inputs as embeddings. Yet, these approaches remain text-based and have not been extended to pathology images.

In this work, we introduce the first VLM-based framework for capturing and aligning pathology visual-language embeddings. Unlike the CLIP-based framework, our approach unifies vision and language embeddings within a single model, significantly improving the capacity to capture cross-modal relationships. We validate the effectiveness of our framework through comprehensive experiments on two tasks: cross-modal retrieval on pathology image-caption datasets and zero-shot patch classification across seven datasets of patch-level pathology images. Our quantitative results indicate that our framework effectively represents multimodal information, achieving superior performance on all tasks compared to CLIP-based models.

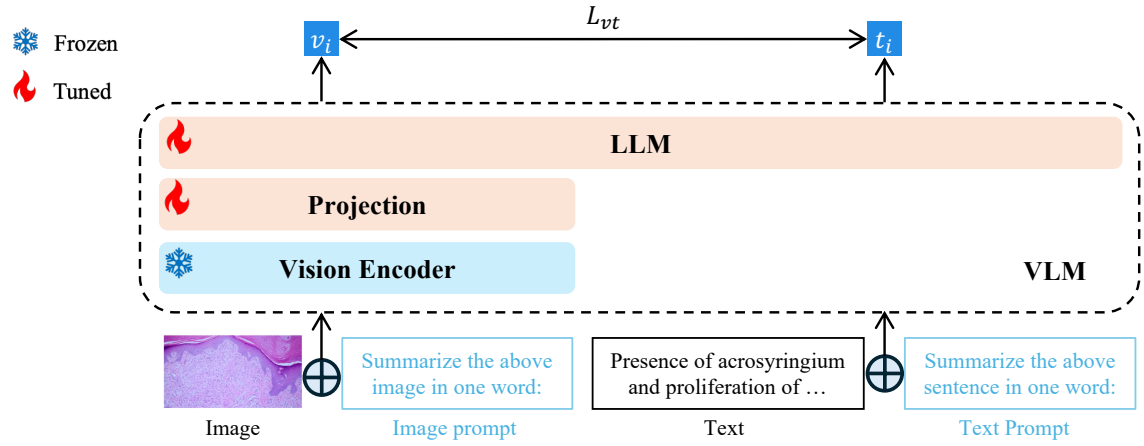

Fig. 1. The architecture of our proposed framework. This framework uses a VLM as the backbone to comprehensively integrate image and text features. By unifying images and texts into the same embedding space with specifically designed prompts, our proposed framework improves multimodal embeddings on image-text pairs using contrastive learning.

2. METHOD

The architecture of our framework is illustrated in Fig. 1. Initially, we design prompts to guide the VLM in generating embeddings (Sec. 2.1). Next, we align vision and language embeddings using contrastive learning (Sec. 2.2). Once pre-training is complete, we transfer the model to perform zero-shot pathology tasks in downstream applications (Sec. 2.3).

2.1. Using VLM to get embeddings

We employ a prompt-based representation method with VLMs inspired by E5-V [8]. The key idea is using specific prompts to instruct VLMs to represent the inputs into words. Given a VLM H and an image-text pair (x, c). We apply the prompt `<image> \n Summarize the above image in one word: <text> \n Summarize the above sentence in one word:` to represent image x and text c to get image input x_{input} and text input c_{input}, where `<image>` and `<text>` are the placeholders for image x and text c. Then we feed this input $(x_{\text{input}}, c_{\text{input}})$ to the VLM H to get the image and text embedding $(\mathbf{v}, \mathbf{t})$, which is the vector of the last token obtained from the final layer of the VLM output.

2.2. Vision-language contrastive learning

Given a batch of N paired image and caption samples $\{(x_n, c_n)\}_{n=1,\ldots,N}$ and a VLM H, for each sample (x_i, c_i), we feed them into VLM H to get the image and text embeddings $(\mathbf{v}_i, \mathbf{t}_i)$ as described in Sec. 2.1. Our goal is to construct a visual-language embedding space from paired image-text data, that satisfies $sim(\mathbf{v}_i, \mathbf{t}_i) \gg sim(\mathbf{v}_i, \mathbf{t}_j)$, $i \neq j$, where the sim denotes the similarity of image embedding $\mathbf{v}$ and text embedding $\mathbf{t}$, we apply $cosine$ function to calculate the similarity. Therefore, we optimize an infoNCE contrastive Loss to train our model, which is $\mathcal{L}_{\text{vt}} = -(\log \frac{exp(cos(\mathbf{v}_i, \mathbf{t}_j)/\tau)}{\sum_{j=1}^{n} exp(cos(\mathbf{v}_i, \mathbf{t}_j)/\tau)} + \log \frac{exp(cos(\mathbf{t}_i, \mathbf{v}_j)/\tau)}{\sum_{j=1}^{n} exp(cos(\mathbf{t}_i, \mathbf{v}_j)/\tau)})$, $i \neq j$, where $\mathbf{v}_i$ and $\mathbf{t}_i$ are the embeddings for the aligned i-th image and text pair, τ denotes a temperature parameter, and $cos(\mathbf{v}_i, \mathbf{t}_i)$ and $cos(\mathbf{t}_i, \mathbf{v}_i)$ denote the two directions of contrastive learning.

2.3. Zero-shot transfer for pathology downstream tasks

The CLIP model [7] introduced a method using prompts for zero-shot classification. In this method, each class name is converted into a sentence by attaching to a specific template. For instance, the class name "Muscularis Mucosa" is expanded using the template "An H&E image of { }" to form the sentence "An H&E image of Muscularis Mucosa." We expand each class name into a sentence using this template. Subsequently, our pre-trained model calculates embeddings for these sentences and for test images, and their similarity is measured as in Sec. 2.2. The labels of the test images are then determined based on the highest similarity scores.

3. EXPERIMENTS

3.1. Experiment settings

We use LLaVA-NeXT-8B [9] as our backbone and fine-tune the checkpoint from E5-V [8] with our training data. All the images are resized to 336 × 336 pixels. We adopt Openpath [3] and Quilt1M [6] as our training data. We evaluate the model on two downstream tasks: cross-modal retrieval and zero-shot patch classification. For cross-modal retrieval task, we use the Recall@K metric. Accuracy and weighted F1 metrics are used for zero-shot patch classification tasks.

Task	Model	Arch-PubMed			Arch-book		
		Recall@1	Recall@10	Recall@50	Recall@1	Recall@10	Recall@50
i2t	PLIP [3]	0.010	0.067	0.185	0.030	0.152	0.393
	QuiltNet [6]	0.018	0.133	0.324	0.041	0.168	0.399
	E5-V [8]	0.004	0.011	0.042	0.003	0.018	0.067
	Ours (Openpath)	0.027	0.144	0.344	0.051	0.214	**0.506**
	Ours (Quilt1M)	**0.036**	**0.182**	**0.399**	**0.064**	**0.244**	0.497
t2i	PLIP [3]	0.008	0.067	0.181	0.033	0.164	0.419
	QuiltNet [6]	0.020	0.114	0.296	0.028	0.204	0.429
	E5-V [8]	0.006	0.022	0.072	0.004	0.026	0.111
	Ours (Openpath)	0.028	0.146	0.349	0.056	0.244	0.541
	Ours (Quilt1M)	**0.039**	**0.164**	**0.409**	**0.064**	**0.297**	**0.584**

Table 1. Performance comparison with PLIP [3], QuiltNet [6] and E5-V [8] on two retrieval datasets. Recall metrics are reported. Task i2t and t2i denote image-to-text and text-to-image retrieval, respectively. Ours (Openpath) and Ours (Quilt1M) denote our model pre-trained on Openpath [3] and Quilt1M [6], respectively. Bold fonts illustrate the best performance.

Model	BACH	CRC-100k	SkinCancer	LC25000	RenalCell	SICAP	WSSS4LUAD
	acc wF1	acc wF1	acc wF1	acc wF1	acc wF1	acc wF1	acc wF1
PLIP [3]	0.266 0.138	0.449 0.420	0.332 0.346	0.655 0.664	0.495 0.419	0.142 0.070	0.443 0.385
QuiltNet [6]	0.441 0.403	0.464 0.415	0.351 0.298	0.401 0.344	0.376 0.382	0.167 0.166	0.396 0.399
E5-V [8]	0.276 0.165	0.248 0.132	0.040 0.016	0.202 0.072	0.159 0.049	0.197 0.174	0.253 0.104
Ours (Openpath)	0.424 0.377	**0.607 0.573**	0.399 0.356	0.653 0.613	0.404 0.416	0.484 0.480	0.459 0.378
Ours (Quilt1M)	**0.516 0.450**	0.477 0.453	**0.412 0.398**	**0.677 0.665**	**0.500 0.493**	**0.599 0.601**	**0.523 0.467**

Table 2. The comparison of zero-shot patch classification between different models. Accuracy (acc) and weight F1 score (wF1) are reported. Bold fonts illustrate the best performance.

3.2. Cross-modal retrieval Results

In Table 1, we demonstrate that our proposed models outperform the baseline models across two different datasets on cross-modal retrieval tasks. For the image-to-text (i2t) retrieval task, our model trained on the Openpath [3] dataset outperforms the PLIP [3] model on both datasets, despite both models using identical training data. Specifically, on the Arch-PubMed dataset, our model achieves a Recall@10 of 0.144, which is more than 7% higher compared with PLIP [3]. Similarly, our Quilt1M [6] pre-trained model outperforms QuiltNet [6] on both datasets. On the Arch-book dataset, it achieves a Recall@50 of 0.497, representing a 10% improvement. A similar trend is observed for the text-to-image (t2i) retrieval task. Our model surpasses PLIP [3] on both datasets, with the most significant improvement seen in Recall@50, reaching 0.541 on the Arch-book dataset compared to PLIP's 0.419. Our Quilt1M [6] pre-trained model further advances this lead by achieving a Recall@50 of 0.584. All these results demonstrate that our framework effectively captures visual language relationships even within the same data constraints. Notably, the E5-V [8] model, which shares the same architecture as our models but lacks fine-tuning on pathology data, exhibits substantially lower recall scores across all tasks and datasets. This difference highlights the importance of fine-tuning with pathology data.

3.3. Zero-shot classification Results

We also compare the zero-shot patch classification performance of different models across eight datasets (Table 2). Our models pre-trained on Openpath [3] and Quilt1M [6] consistently outperform the baseline models, highlighting the benefits of our framework and the importance of domain-specific pretraining. For example, on the BACH [10] dataset, our model pre-trained on Quilt1M [6] achieves the highest accuracy and wF1 (0.516 and 0.450, respectively), exceeding the CLIP-based QuiltNet [6] model by over 7% in accuracy. Similarly, on the CRC-100K [11] dataset, the Openpath model achieves the highest accuracy and wF1 (0.607 and 0.573), significantly outperforming PLIP [3], which also utilizes the Openpath [3] dataset. These results highlight our model's superior feature extraction and classification capabilities within the same data constraints. On other datasets, such as SkinCancer [12] and RenalCell [13], our model achieves consistently superior performance, demonstrating high adaptability across various pathology domains. Notably, on the LC25000 [14] dataset, our model trained on Quilt1M [6] achieves a wF1 of 0.665, approximately 60% higher than E5-V [8], a VLM-based model without pathology-specific fine-tuning. This finding emphasizes the importance of training with pathology-relevant data, further validating the effectiveness of our framework.

4. CONCLUSION

This study presents a VLM-based framework specifically designed for capturing and aligning pathology visual-language embeddings. Unlike CLIP-based methods, our framework uses a unified model to integrate vision and language modalities, enhancing cross-modal relational understanding. Extensive evaluations on cross-modal retrieval and zero-shot classification tasks demonstrate that our framework outperforms CLIP-based models. Our study provides a foundation for future exploration of VLM applications in computational pathology, emphasizing the role of advanced multimodal techniques in improving diagnostic and research capabilities.

5. ACKNOWLEDGEMENTS

This work was partially supported by CPRIT RP230363, NSF CCF-2400785 and NSF IIS-2412195.

6. COMPLIANCE WITH ETHICAL STANDARDS

Ethical approval was not required as confirmed by the license attached with the open-access data. This study was performed in line with the principles of the Declaration of Helsinki.

7. REFERENCES

[1] Qifeng Zhou, Wenliang Zhong, Yuzhi Guo, Michael Xiao, Hehuan Ma, and Junzhou Huang, "Pathm3: A multimodal multi-task multiple instance learning framework for whole slide image classification and captioning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 373–383.

[2] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Medical Image Analysis*, vol. 65, pp. 101789, 2020.

[3] 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*, vol. 29, no. 9, pp. 2307–2316, 2023.

[4] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang, "Wsisa: Making survival prediction from whole slide histopathological images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7234–7242.

[5] 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*, 2021, pp. 16549–16559.

[6] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro, "Quilt-1m: One million image-text pairs for histopathology," *Advances in neural information processing systems*, vol. 36, 2024.

[7] 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*. PMLR, 2021, pp. 8748–8763.

[8] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang, "E5-v: Universal embeddings with multimodal large language models," *arXiv preprint arXiv:2407.12580*, 2024.

[9] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li, "Llava-next: Stronger llms supercharge multimodal capabilities in the wild," 2024.

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

[11] Jakob Nikolas Kather, Niels Halama, and Alexander Marx, "100,000 histological images of human colorectal cancer and healthy tissue," May 2018.

[12] Katharina Kriegsmann, Frithjof Lobers, Christiane Zgorzelski, Jörg Kriegsmann, Charlotte Janßen, Rolf Rüdinger Meliß, Thomas Muley, Ulrich Sack, Georg Steinbuss, and Mark Kriegsmann, "Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections," *Frontiers in Oncology*, vol. 12, pp. 1022967, 2022.

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

[14] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides, "Lung and colon cancer histopathological image dataset (lc25000)," *arXiv preprint arXiv:1912.12142*, 2019.