Advances in Artificial Intelligence and Machine Learning; Research (Ahead of Print) Received 27-01-2026; Accepted 02-05-2026; Published online 09-05-2026

Content-Based Image Retrieval in Histopathology and Gastrointestinal Endoscopy: A Comparative Study of Deep Models

Ciprian-Mihai Ceausescu ciprian-mihai.ceausescu@drd.unibuc.ro
Faculty of Mathematics and Computer Science
University of Bucharest

Corresponding Author: Ciprian-Mihai Ceausescu

Abstract

In the medical domain, finding the right diagnosis rarely happens using just one piece of information. The specialists have to put together several indicators to provide an answer to a specific patient. Content-based image retrieval is becoming a relevant task in the medical image analysis field. It allows the doctors to search within large medical image databases for cases that look visually similar, pull those cases up, and directly compare them with the current patient's record. In this work, we propose a comparison protocol between several pretrained backbone models on three different medical imaging datasets: lung histopathology, colon histopathology, and wireless capsule endoscopy videos. The models we take into account include classic convolutional networks, newer transformer-based architectures, self-supervised approaches, multimodal encoders, and even some models originally built for segmentation tasks. We propose an evaluation framework across diverse medical domains (from tiny microscope slides of tissue to wide-angle endoscopic views inside the gut). The image processing and embedding retrieval are performed using the same pipeline, gaining a better overview of how well different pretrained representations transfer between these domains. From a methodological point of view, our study tries to fix a common gap in the literature: a consistent comparison of backbones for medical retrieval on histopathology and gastrointestinal datasets. From a practical point of view, the results should help specialists decide which models tend to work best for medical image search tasks. Our findings provide valuable insights when building retrieval-based tools to support diagnosis, train new doctors, or simply browse and understand large clinical datasets. Overall, the benchmark can serve as a starting point for future work on medical image retrieval and for adapting large pretrained models to different clinical imaging domains.

Keywords: Content-based image retrieval, Medical image analysis, Histopathology, Wireless capsule endoscopy, Foundation models, Image embeddings, Benchmark, Deep learning.

1. INTRODUCTION

Content-based image retrieval (CBIR) methods started to gain interest in the medical image analysis field, especially when predicting a label for a specific piece of information is not

Citation: Ciprian-Mihai Ceausescu Content-Based Image Retrieval in Histopathology and Gastrointestinal Endoscopy: A Comparative Study of Deep Models. **Advances in Artificial Intelligence and Machine Learning**.2026. (Ahead of Print) https://dx.doi.org/10.54364/AAIML.2026.62301

enough. In several medical applications, it is useful to retrieve similar examples from a large database of archived records and to analyze them, rather than making a decision based only on the patient's data. This time-consuming process makes the data retrieval process useful for tasks that depend strongly on visual comparison and interpretation. In practice, CBIR methods can support case-based reasoning and can help clinicians inspect related examples, making model behavior easier to interpret, providing the patient with a better diagnosis.

CBIR methods are relevant in areas such as histopathology and gastrointestinal endoscopy, where specialists work with a high amount of data. In histopathology, diagnosis often depends on fine-grained visual cues, including tissue organization, glandular structure, and nuclear appearance. Deep learning has already shown strong results for classification and segmentation in this domain [1, 2], but retrieval offers something different. Image retrieval task can allow direct comparison between new samples and the existing samples that are stored in the hospital databases. Therefore, retrieving similar images and then processing them can be useful in several situations. The doctors can use similar cases in the diagnosis process of a new patient or to educate the young specialists to find unusual, rare cases that can be analyzed. In gastrointestinal endoscopy, where physicians often rely on texture, color, and local visual context to identify abnormalities such as inflammation, polyps, or mucosal lesions [3], CBIR methods can be widely used.

Meanwhile, visual representation learning of the data has advanced rapidly, with the newly developed methods being capable of obtaining high-quality results. CNN models, such as ResNet [4], established strong baselines for image understanding. More recently, transformer-based architectures, including ViT [5], and Swin Transformer [6], have shown a strong ability to capture both the global structures of the images and their local details. Furthermore, self-supervised approaches, such as DINOv2 [7] and MAE [8], showed that meaningful image embeddings can be learned even without dense manual annotation of the training data.

Latest advances in the field have expanded this even more by allowing the models to be trained using two encoders, one specialized for text and one specialized for images. Contrastive Language-Image Pre-training (CLIP) [9], introduced multimodal contrastive learning by aligning image and text embeddings in a shared space. SigLIP [10], refined this paradigm by replacing the standard contrastive loss with a sigmoid-based objective. Following this approach, SigLIP introduces more stable training and stronger image-text alignment. OpenCLIP [11], extended this idea by training large-scale variants. ViT-H/14 model was trained on broad open-source datasets, providing powerful general-purpose visual encoders. Domain-specific extensions have also proven effective. The BiomedCLIP [12], model adapted contrastive image-text learning to biomedical data, and MedSigLIP was built on the SigLIP framework to target medical imaging specifically. More recently, together with these multimodal approaches, UNI [13], represents a dedicated pathology foundation model. It was trained via self-supervised learning on large-scale histopathology data, offering representations that are particularly fine-tuned to tissue-level visual patterns. Large pretrained encoders from segmentation-oriented foundation models, including SAM [14] and MedSAM [15], also provide rich visual features that may be useful beyond segmentation itself.

Even with all of this progress, it is still not very clear which pretrained models work best for medical image retrieval. Most prior studies emphasize classification or segmentation performance, but retrieval depends on different features of the data. Strong retrieved em-

beddings should keep visually similar cases close together while separating different classes in a meaningful way. This is not necessarily reflected by classification accuracy alone. In addition, the behavior of pretrained representations can change a lot across medical domains because the visual characteristics of microscopic tissue images and endoscopic images are quite different. Although recent work has started to examine pretrained models for medical CBIR [16], broader comparisons across distinct domains are still relatively limited.

In this study, we benchmark a diverse set of pretrained backbone models for CBIR on three datasets drawn from two medical imaging domains: histopathology and wireless capsule endoscopy. The histopathology dataset contains five tissue classes and is divided into lung and colon subsets, which we treat as two separate retrieval tasks. The third dataset is a curated WCE dataset with four gastrointestinal classes. To keep the comparison fair, all backbones are evaluated under the same retrieval pipeline based on frozen feature extraction, cosine similarity, and nearest-neighbor search.

The main contribution of this work is a consistent comparison of several backbone families under the same evaluation setting. These include CNN-based models, transformer architectures, self-supervised encoders, multimodal models, among them SigLIP, MedSigLIP, and OpenCLIP ViT-H/14, the pathology foundation model UNI, and segmentation-oriented backbones. We also study how retrieval performance changes across imaging modalities, which gives some insight into the effect of domain shift. Finally, we provide a reproducible benchmarking pipeline that can be extended to other datasets and additional pretrained models. Overall, the goal is to offer a practical reference point for selecting feature extractors in medical image retrieval and to provide a baseline for future work in this area.

2. RELATED WORK

2.1 Medical Image Retrieval

Content-based image retrieval has a long history in medical imaging, where early approaches relied on handcrafted features such as texture descriptors, color histograms, and local binary patterns [17–19]. These methods were limited in their ability to capture complex visual patterns present in medical data. With the recent developments in the deep learning domain, the feature extraction process has shifted toward learning representations from convolutional neural networks, which significantly improved the retrieval performance [2].

Some recent studies have explored deep CBIR systems for histopathology and gastrointestinal imaging. In histopathology, deep features have been used to retrieve similar tissue patterns and assist in cancer diagnosis [1]. The developed CBIR systems have also been applied to endoscopic images, to distinguish between lesions and abnormal mucosal structures [3]. Despite the advances in the field, most articles focus on a single architecture or dataset design. Therefore, the comparison of different backbone families under consistent conditions is more difficult.

2.2 Representation Learning for Retrieval

The recently developed retrieval systems are based on representation learning. Therefore, CNN-based models such as ResNet [4], have been widely used due to their strong performance and computational efficiency. Transformer-based architectures, including ViT [5] and Swin Transformer [6], provide improved global context modeling and have shown competitive or superior results in many vision tasks.

In the medical domain setup, where acquiring data is a time-consuming task, self-supervised learning methods have gained attention, mainly because their training process does not require labeled data. Models such as DINOv2 [7] and MAE [8], are designed to learn general-purpose visual features that transfer well across tasks, including content-based image retrieval. These approaches are particularly relevant in the medical imaging field, where annotated data can be scarce.

Recently, multimodal architectures that are trained via contrastive image-text objectives have demonstrated strong zero-shot transfer capabilities by learning a shared embedding space. This approach will map close together the semantically related images and their text description, empowering the model to recognize or retrieve concepts not seen before in a downstream task. CLIP [9], introduced this paradigm, aligning images and texts using two different encoders. SigLIP [10], further refined this approach by replacing the global softmax loss with a pairwise sigmoid objective, which improves the training efficiency and the retrieval alignment. OpenCLIP [11], extended this paradigm even more by training large-scale architectures, including a ViT-H/14 model, on open-source datasets. Thus, powerful general-purpose visual encoders are developed. Following the same direction, domain-specific adaptations are proposed. In the BiomedCLIP [12], architecture, biomedical image-text pairs are used, optimizing and capturing medical semantics, while MedSigLIP[20], applies the sigmoid loss framework specifically to medical imaging data. These models have shown promise in retrieval settings, especially when semantic alignment between image content and clinical concepts is important.

2.3 Foundation Models in Medical Imaging

Large-scale foundation models have recently emerged as a new paradigm in computer vision. The Segment Anything Model (SAM) [14], is trained on a massive dataset for segmentation tasks and produces rich image embeddings. Domain-specific adaptations, such as MedSAM [15], further tailor these representations to medical images. Although these models were not originally designed for retrieval, their embeddings may contain useful structural information for similarity-based search.

A complementary direction has emerged in the form of pathology-specific foundation models. UNI [13], is a self-supervised encoder pretrained on over 100 million histopathology image patches spanning more than 100.000 whole-slide images across 20 tissue types. Unlike general-purpose encoders, UNI is designed to capture the fine-grained morphological detail characteristic of tissue images, making it a particularly relevant backbone for histopathology retrieval tasks.

The effectiveness of these diverse model families for CBIR nonetheless remains an open question, especially when compared under consistent evaluation conditions and across imaging modalities with very different visual characteristics. Overall, there is a lack of comprehensive benchmarks that compare diverse backbone architectures for medical image retrieval across multiple domains. This study addresses this gap by providing a unified evaluation framework.

3. MATERIALS AND METHODS

3.1 Datasets

In our work, the experimental evaluation of the image retrieval pipeline is conducted on three medical imaging datasets spanning two distinct visual domains: (i) histopathology, and (ii) wireless capsule endoscopy (WCE). We aim to assess whether pretrained visual backbones produce discriminative embeddings that remain effective across datasets with different appearance statistics, structural patterns, and acquisition conditions.

3.1.1 Histopathology source dataset

The histopathology dataset that we used in our study was derived from the LC25000 collection [21]. This dataset contains around 25.000 RGB images, having 768×768 pixels as spatial resolution for each image. The dataset is a smaller validated subset that contains both lung and colon tissue images. The dataset was further expanded by augmenting the existing data to increase the generalization of the developed methods. In the dataset, we can find five balanced diagnostic categories, each of them having 5.000 images. The lung dataset partition has the following classes: (i) lung benign tissue; (ii) lung adenocarcinoma tissue; and (iii) lung squamous cell carcinoma. Meanwhile, the colon dataset partition has the following classes: (i) colon benign tissue; and (ii) colon adenocarcinoma. Our aim is to examine the retrieval process and how this varies under two different classification tasks. Each of the two subsets has a different level of ambiguity and difficulty, making it clinically relevant for two different medical specializations. For the lung dataset, having a three-class configuration helps the retrieval system to focus on different malignant tissue and to distinguish against the benign tissue. Therefore, we can further analyze the similarities between different classes. Usually, the malignant images are very similar when they are visualized with the microscope, and the image retrieval task becomes more challenging. For the colon tissue images, the task is simplified by the structure of the dataset (two-class setup instead of three). The feature extractors focus on classifying between benign and malignant images. Comparing how the retrieval systems work in both scenarios will help the specialists to find the best backbone depending on the given task.

3.1.2 Wireless capsule endoscopy dataset

The third dataset that we analyze in our work is a curated WCE colon dataset. It contains four different classes, namely: (i) normal; (ii) ulcerative colitis; (iii) polyps; (iv) esophagitis. Compared to the histopathology dataset, where the images contain cellular structures, the WCE colon dataset is variant in illuminance level, viewpoint, and mucosal texture. The existing features make the task more challenging, having domain-shifted data.

Including WCE in the benchmark makes it possible to examine whether the same pretrained backbone remains robust across both microscopic and endoscopic image domains. In the broader gastrointestinal imaging literature, curated WCE datasets with disease-oriented labels have been used to study automated analysis, benchmarking, and dataset standardization [3].

3.2 Study Design

Each existing dataset that we analyze in our work was evaluated in terms of retrieval performance using the same protocol, allowing the backbones to be compared under the same experimental conditions. The images are merged together, and the whole dataset is split into 80% used to construct the embedding database and the remaining 20% used for the query set.

For each query image that we analyze from the query set, we extract feature embeddings using different frozen pretrained models. These embeddings are then compared with all the embeddings from the database, constructed before from all the images from the training split. The embedding items were ranked according to cosine similarity, and the top-k most similar images are returned. To evaluate the retrieval procedure, we compare the class label of the query image with the labels of the top-k most similar images. In FIGURE 1, we present the complete pipeline of our proposed retrieval method. The two steps (indexing step and retrieval step) are presented for the WCE dataset, but this pipeline is generally defined by changing the dataset or by merging all the datasets together. We can observe how the embedding database is generated and how the retrieval process returns the more similar images.

Our benchmark was designed to answer three central questions: (i) first, we asked how well different pretrained backbones perform as frozen feature extractors for medical image retrieval without any task-specific fine-tuning; (ii) second, we examined whether the strongest models remain consistently strong across both histopathology and endoscopy; (iii) third, we investigated which families of pretrained models are more robust to domain changes involving texture, structure, scale, and acquisition style.

3.3 Backbone Architectures

We evaluated a diverse set of convolutional, transformer-based, self-supervised, multimodal, and segmentation-oriented backbones: ResNet50, ViT-B/16, Swin-B, CLIP ViT-L/14, Biomed-

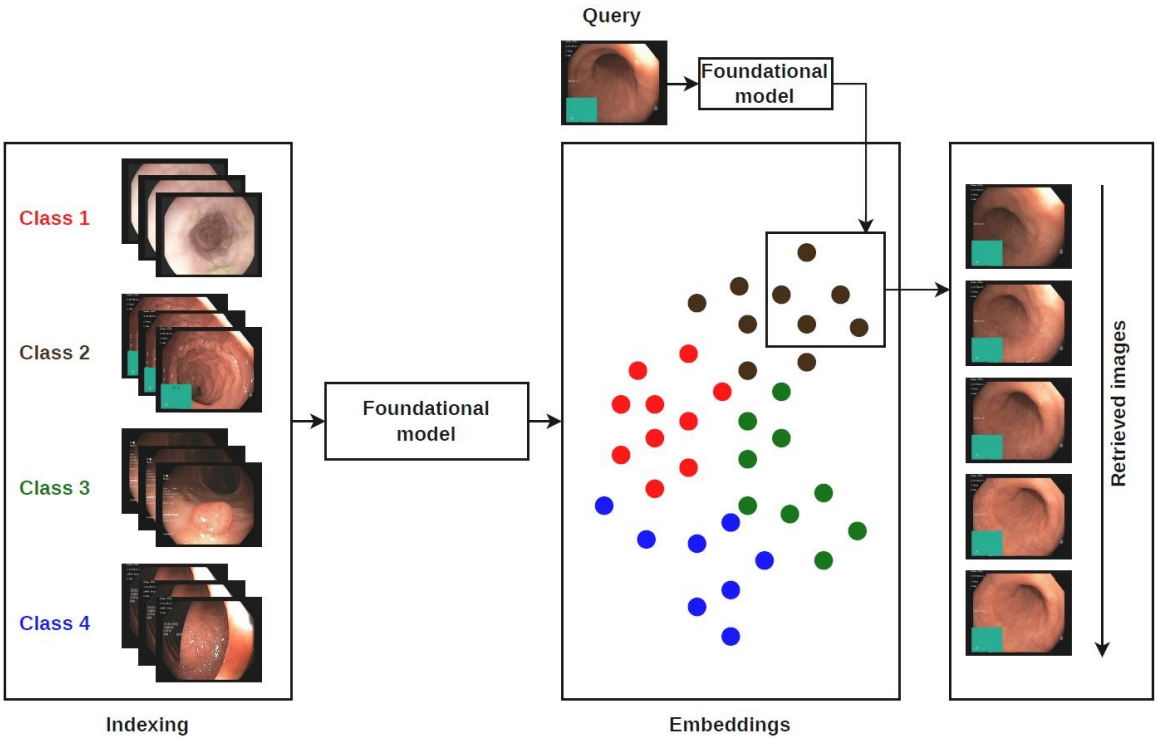


Figure 1: Overview of the image retrieval pipeline. Input images are organized into class-specific folders, and feature embeddings are extracted using a frozen pretrained backbone. Query embeddings obtained using the same frozen pretrained backbone are compared to stored embeddings, and the top-k most similar images are retrieved.

CLIP, SigLIP, MedSigLIP, OpenCLIP ViT-H/14, DINOv2 ViT-B/14, MAE ViT-B/16, UNI, BMC-CLIP-CF, SAM ViT-B, and MedSAM ViT-B.

These models that we use in our work cover different pretraining strategies. ResNet50 [4], represents a powerful convolutional neural network baseline, having a good transfer knowledge (between different imaging modalities) performance. ViT-B/16 [5] (patch-based self-attention), and Swin-B [6] (hierarchical, multi-scale attention), are transformer-based visual encoders that were trained for general recognition tasks.

CLIP ViT-L/14 [9], BiomedCLIP [12], BMC-CLIP-CF [22], SigLIP [10], MedSigLIP, and OpenCLIP ViT-H/14 [11], were selected to represent contrastive vision-language pretraining. These models contain a contrastive loss function, encouraging the embeddings that are extracted to be semantically relevant.

SigLIP and MedSigLIP represent an extension of these models. They are replacing the standard softmax loss with a pairwise sigmoid loss. OpenCLIP ViT-H/14 architecture defines a general-purpose encoder that is trained at scale on open-source data. Therefore, this model provides a higher generalization.

DINOv2 [7] and MAE [8], represent self-supervised methods, trained to capture strong transferable visual structures, without the need for pair text supervision training. UNI [13], model extends this paradigm towards histopathology domains. It provides a self-supervised extractor that was trained on a large amount of histopathology data, mainly for tissue-level retrieval problems. Finally, the last two backbones used in our work, SAM [14], and MedSAM [15], were evaluated as segmentation encoders. Their representations can be useful for similarity-based retrieval, due to the powerful encoders trained to capture fine-grained details.

3.4 Retrieval Pipeline

All datasets used in the present work were organized in a class-wise directory structure, where each class has a semantic meaning associated with its images.

Prior to feature extraction, all images were uniformly preprocessed, including resizing to match the input resolution required by each backbone and applying standard normalization procedures. Most models operated on inputs of size 224×224, while architectures based on Segment Anything employed higher resolutions, such as 512×512 or 1024×1024.

To enable fair comparison across different visual encoders, we designed a unified feature extraction framework. Regardless of architectural differences, each model produced a single global descriptor per image. For models generating spatial or token-level representations, global pooling was applied to obtain fixed-length feature vectors. In some cases, feature representations were projected into a common embedding space to ensure consistency across encoders. Finally, all embeddings were normalized using $L2$ normalization.

3.5 Implementation Details

Our experimental configuration was created to include several pretrained models (on natural or medical images) through a single evaluation framework. Thus, the experimental parameters were the same for all the evaluated backbones. We set the batch size to 32, the number of data-loading workers to 2, and the target embedding dimension after optional projection to 512. All the visual encoders were used in evaluation mode, and gradient computation was disabled throughout embedding extraction to ensure that no task-specific adaptation was performed.

The implementation accounted for practical differences between model families. Standard CNN and transformer models from `timm` were used through their `forward_features` interface, which returns pooled representations without the classification head. CLIP-based models used their vision encoder directly via `CLIPVisionModel`, while OpenCLIP-based models, including BiomedCLIP, BMC-CLIP-CF, and OpenCLIP ViT-H/14, used the `encode_image` interface exposed by `open_clip` library. SigLIP and MedSigLIP were loaded via `AutoModel` from HuggingFace and queried through their `get_image_features` method when available, falling back to mean-pooling over the last hidden state otherwise. MedSigLIP additionally required a larger input resolution of 448×448 pixels to match its pretraining configuration, while all other models operated at 224×224. UNI was loaded as a `timm` model using the

HuggingFace Hub path and queried through `forward_features`, with additional initialization arguments (`init_values` and `dynamic_img_size`) applied when required by the model card. DINOv2 used the corresponding `forward_features` interface provided by the Facebook Research hub model. For SAM and MedSAM, image embeddings were obtained from the image encoder branch and pooled into global descriptors; since these encoders are memory-intensive, inference was performed in smaller chunks when needed. All embeddings were L2-normalized before similarity computation, regardless of the backbone used.

3.6 Evaluation Metrics

Retrieval performance was evaluated using mean Average Precision (mAP), Mean Reciprocal Rank (MRR), and Precision@K with $K \in \{1, 5, 10\}$. These metrics were computed by comparing the class label of each query image with the labels of the retrieved gallery samples.

Let q denote a query image, and let the retrieved gallery images be ranked in descending order according to cosine similarity. Let $rel_i \in \{0, 1\}$ indicate whether the retrieved image at rank i belongs to the same class as the query. Precision@K is defined as

$$\text{Precision@K} = \frac{1}{K} \sum_{i=1}^{K} rel_i. \tag{1}$$

Average Precision (AP) summarizes the ranking quality for a single query:

$$\text{AP}(q) = \frac{1}{R_q} \sum_{i=1}^{N} P(i) \cdot rel_i, \tag{2}$$

where $P(i)$ denotes precision at rank i, N is the number of retrieved items, and R_q is the total number of relevant gallery samples for query q. The final mAP score is obtained by averaging AP over all queries.

Mean Reciprocal Rank (MRR) evaluates how early the first correct retrieval appears:

$$\text{MRR} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{rank_q}, \tag{3}$$

where $rank_q$ represents the position of the first relevant image that was retrieved (from the training dataset) for query image q, and Q is the total number of queries.

These metrics, used together, offer a good overview of the pretrained backbones that we use in our retrieval pipeline.

4. RESULTS

4.1 Overall Retrieval Performance

In the TABLE 1, we summarize the results of the retrieval pipeline of all the proposed backbones on the three datasets (lung histopathology, colon histopathology, and WCE). Initially, our results are reported separately for the lung histopathology benchmark, the colon histopathology benchmark, and the WCE benchmark.

Across datasets, mAP serves as the main global indicator of ranking quality, while MRR reflects how rapidly the first correct match is retrieved. Precision@1 is especially relevant in medical retrieval because it measures whether the top returned example is already diagnostically or visually meaningful.

4.2 Lung Histopathology Retrieval

The initial benchmark includes the Lung Histopathology datasets and evaluates the instance retrieval method for the following dataset classes: (i) lung_aca; (ii) lung_n; and (iii) lung_scc. This subset contains a total of 15.000 images (12.000 training and 3.000 testing examples). In this setup, the retrieval pipeline focused on tissue structures, cells, and stain properties. Errors are more likely due to intra-class variability, similar textures across diseases, and differences in staining or patch composition, rather than viewpoint or acquisition artifacts. This makes Lung Histo a suitable benchmark for evaluating fine-grained morphological representation quality.

On Lung Histo, overall performance is extremely high across most methods, with the majority of backbones achieving near-saturated retrieval scores. The best mAP is achieved by **UNI**, which reaches **0.9999**, together with a perfect MRR of **1.0000**, P@1 of **1.0000**, P@5 of **0.9999**, and P@10 of **0.9996**. This result is consistent with UNI being a self-supervised pathology foundation model pretrained on large-scale histopathology data, and confirms that domain-specific pretraining confers a clear advantage when the retrieval task closely matches the pretraining distribution. **CLIP ViT-L/14** is a very close second with mAP **0.9988**, MRR **0.9998**, and P@1 **0.9997**, followed by **Swin-B**, which attains the joint-best exact top-rank performance with P@1 of **1.0000** and MRR of **1.0000**, alongside an mAP of **0.9987**. **BMC-CLIP-CF** also performs exceptionally well at mAP **0.9980** and P@1 **0.9997**. Among the remaining top-tier models, **DINOv2 ViT-B/14**, **ResNet50**, and **BiomedCLIP** all achieve near-saturated retrieval performance with mAP values of **0.9976**, **0.9976**, and **0.9970**, respectively. **SigLIP** and **OpenCLIP ViT-H/14** occupy a solid intermediate tier, reaching mAP values of **0.9943** and **0.9924**, respectively. Both models remain strong on this task, though they fall slightly below the top contrastive and domain-specific encoders. **MedSigLIP** performs much lower than the SigLIP backbone, having a mAP value of **0.9729**. Thus, it suggests that medical adaptation using the sigmoid contrastive loss function doesn't improve histopathology retrieval.

MAE ViT-B/16 and **ViT-B/16** achieve mAP values of **0.9878** and **0.9834**. They represent competitive methods, but consistently overperformed by the top models across all metrics.

Table 1: Comparison of retrieval performance of various backbone models across three medical imaging datasets (Lung Histo, Colon Histo, and WCE). Metrics include mean Average Precision (mAP), Mean Reciprocal Rank (MRR), and Precision at top 1, 5, and 10 results (P@1, P@5, P@10). Best-performing results for each dataset and metric are highlighted in bold.

Backbone	Dataset	mAP	MRR	P@1	P@5	P@10
ResNet50	Lung Histo	0.9976	0.9996	0.9993	0.9969	0.9886
ViT-B/16	Lung Histo	0.9834	0.9939	0.9900	0.9756	0.9589
Swin-B	Lung Histo	0.9987	**1.0000**	**1.0000**	0.9987	0.9930
CLIP ViT-L/14	Lung Histo	0.9988	0.9998	0.9997	0.9985	0.9945
BiomedCLIP	Lung Histo	0.9970	0.9995	0.9993	0.9959	0.9884
DINOv2 ViT-B/14	Lung Histo	0.9976	0.9993	0.9990	0.9967	0.9902
MAE ViT-B/16	Lung Histo	0.9878	0.9975	0.9957	0.9820	0.9630
BMC-CLIP-CF	Lung Histo	0.9980	0.9998	0.9997	0.9970	0.9917
SAM ViT-B	Lung Histo	0.9685	0.9893	0.9820	0.9530	0.9298
MedSAM ViT-B	Lung Histo	0.6673	0.7390	0.5860	0.5782	0.5748
SigLIP	Lung Histo	0.9943	0.9993	0.9990	0.9924	0.9822
MedSigLIP	Lung Histo	0.9729	0.9894	0.9833	0.9611	0.9450
UNI	Lung Histo	**0.9999**	**1.0000**	**1.0000**	**0.9999**	**0.9996**
OpenCLIP ViT-H/14	Lung Histo	0.9924	0.9990	0.9983	0.9894	0.9759
ResNet50	Colon Histo	0.9994	0.9998	0.9995	0.9992	0.9984
ViT-B/16	Colon Histo	0.9885	0.9962	0.9940	0.9835	0.9713
Swin-B	Colon Histo	0.9999	**1.0000**	**1.0000**	**1.0000**	0.9994
CLIP ViT-L/14	Colon Histo	0.9999	**1.0000**	**1.0000**	0.9999	0.9989
BiomedCLIP	Colon Histo	0.9998	**1.0000**	**1.0000**	0.9997	0.9992
DINOv2 ViT-B/14	Colon Histo	0.9999	**1.0000**	**1.0000**	0.9998	0.9988
MAE ViT-B/16	Colon Histo	0.9978	**1.0000**	**1.0000**	0.9966	0.9907
BMC-CLIP-CF	Colon Histo	0.9998	**1.0000**	**1.0000**	0.9999	0.9990
SAM ViT-B	Colon Histo	0.9706	0.9895	0.9840	0.9585	0.9340
MedSAM ViT-B	Colon Histo	0.6301	0.7059	0.5230	0.5339	0.5306
SigLIP	Colon Histo	0.9821	0.9954	0.9925	0.9750	0.9579
MedSigLIP	Colon Histo	0.9226	0.9568	0.9335	0.8923	0.8729
UNI	Colon Histo	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
OpenCLIP ViT-H/14	Colon Histo	0.9977	**1.0000**	**1.0000**	0.9969	0.9922
ResNet50	WCE	0.9573	0.9791	0.9675	0.9385	0.9268
ViT-B/16	WCE	0.9486	0.9793	0.9692	0.9248	0.9063
Swin-B	WCE	0.9684	0.9901	0.9850	0.9505	0.9372
CLIP ViT-L/14	WCE	0.9644	0.9874	0.9800	0.9458	0.9313
BiomedCLIP	WCE	0.9668	0.9803	0.9717	0.9515	0.9413
DINOv2 ViT-B/14	WCE	0.9565	0.9873	0.9800	0.9347	0.9189
MAE ViT-B/16	WCE	0.9131	0.9484	0.9200	0.8797	0.8560
BMC-CLIP-CF	WCE	**0.9809**	**0.9927**	**0.9892**	**0.9697**	**0.9620**
SAM ViT-B	WCE	0.9207	0.9464	0.9225	0.8932	0.8782
MedSAM ViT-B	WCE	0.5512	0.6094	0.4325	0.4403	0.4396
SigLIP	WCE	0.9203	0.9594	0.9383	0.8818	0.8637
MedSigLIP	WCE	0.8935	0.9368	0.9033	0.8473	0.8168
UNI	WCE	0.9521	0.9741	0.9617	0.9330	0.9146
OpenCLIP ViT-H/14	WCE	0.9698	0.9859	0.9775	0.9548	0.9449

Standard transformer representations and masking procedure are effective for histopathology retrieval, but do not fully match the top-performing encoders.

A clearer performance decline is observed for segmentation-oriented models. **SAM ViT-B** reaches **0.9685** mAP, which remains respectable but is notably below the leading backbones. The most pronounced degradation is seen for **MedSAM ViT-B**, which performs far worse than all other models with only **0.6673** mAP, **0.7390** MRR, and **0.5860** P@1. This gap shows

Table 2: Comparison of retrieval performance of various backbone models on the merged medical imaging datasets (Lung Histo, Colon Histo, and WCE). Metrics include mean Average Precision (mAP), Mean Reciprocal Rank (MRR), and Precision at top 1, 5, and 10 results (P@1, P@5, P@10). Best-performing results for each metric are highlighted in bold.

Backbone	Dataset	mAP	MRR	P@1	P@5	P@10
ResNet50	Merged datasets	0.9896	0.9953	0.9927	0.9855	0.9780
ViT-B/16	Merged datasets	0.9781	0.9939	0.9903	0.9690	0.9530
Swin-B	Merged datasets	0.9920	0.9966	0.9948	0.9885	0.9834
CLIP ViT-L/14	Merged datasets	0.9928	0.9975	0.9961	0.9887	0.9841
BiomedCLIP	Merged datasets	0.9921	0.9961	0.9948	0.9883	0.9823
DINOv2 ViT-B/14	Merged datasets	0.9897	0.9960	0.9942	0.9843	0.9770
MAE ViT-B/16	Merged datasets	0.9750	0.9868	0.9800	0.9645	0.9513
BMC-CLIP-CF	Merged datasets	**0.9947**	**0.9979**	**0.9969**	**0.9926**	**0.9884**
SAM ViT-B	Merged datasets	0.9600	0.9813	0.9715	0.9421	0.9179
MedSAM ViT-B	Merged datasets	0.5377	0.6147	0.4260	0.4153	0.4034
SigLIP	Merged datasets	0.9768	0.9910	0.9861	0.9665	0.9526
MedSigLIP	Merged datasets	0.9376	0.9676	0.9511	0.9120	0.8928
UNI	Merged datasets	0.9902	0.9945	0.9919	0.9866	0.9830
OpenCLIP ViT-H/14	Merged datasets	0.9902	0.9972	0.9956	0.9854	0.9767

segmentation-based models produce weaker embeddings for fine-grained histopathology retrieval, where morphology-aware similarity matters more.

Overall, the Lung Histo results confirm that retrieval performance is nearly saturated for the strongest backbones, with **UNI**, **CLIP ViT-L/14**, and **Swin-B** emerging as the top-performing models. The strong result of UNI in particular underscores the value of domain-aligned pretraining for histopathology retrieval. **ResNet50**, **DINOv2 ViT-B/14**, **BiomedCLIP**, and **BMC-CLIP-CF** also remain highly competitive, confirming that both convolutional and transformer-based encoders pretrained at scale represent histopathological similarity effectively. In contrast, models designed primarily for reconstruction or segmentation show reduced transferability, especially **MedSAM ViT-B**, while sigmoid-based contrastive models, including **SigLIP** and **MedSigLIP**, perform well but do not reach the level of the top contrastive encoders on this domain.

4.3 Colon Histopathology Retrieval

The Colon Histo benchmark evaluates retrieval on histopathology patches acquired under microscopy, where discriminative evidence is driven primarily by glandular architecture, cellular organization, nuclear morphology, staining appearance, and local tissue texture. Retrieval errors in this setting are more likely to arise from subtle intra-class variability, patch-level resemblance between normal and malignant regions, and staining fluctuations rather than from viewpoint or illumination artifacts typical of endoscopy. This dataset provides a useful benchmark for fine-grained morphological retrieval under highly structured visual conditions.

On Colon Histo, the most remarkable result is achieved by **UNI**, which attains a perfect mAP of **1.0000** with flawless scores across every metric, including P@1, P@5, and P@10 all equal to **1.0000**. This is the only instance of perfect retrieval performance across all backbones and datasets in this benchmark, and directly reflects the advantage of domain-specific self-supervised pretraining on large-scale histopathology data when the retrieval task closely matches the pretraining distribution. A cluster of models follows with near-saturated performance: **Swin-B** and **CLIP ViT-L/14** both reach mAP **0.9999** with perfect P@1 and MRR, alongside **DINOv2 ViT-B/14** (mAP **0.9999**), **BiomedCLIP** (mAP **0.9998**), and **BMC-CLIP-CF** (mAP **0.9998**), all effectively achieving perfect top-1 retrieval. **ResNet50** also remains extremely competitive at mAP **0.9994** and P@1 **0.9995**, showing that a strong convolutional baseline is sufficient for this well-separated two-class setting.

OpenCLIP ViT-H/14 performs well at mAP **0.9977** with perfect P@1, though it trails the top contrastive and domain-specific encoders slightly in ranking consistency at deeper positions. **MAE ViT-B/16** similarly achieves perfect P@1 but shows a small degradation in P@5 and P@10 (mAP **0.9978**), suggesting minor inconsistencies in ranking beyond the top result. **ViT-B/16** performs lower than the leading group at mAP **0.9885** and P@1 **0.9940**.

A more substantial drop is visible for the sigmoid contrastive models. **SigLIP** reaches mAP **0.9821** and P@1 **0.9925**, while **MedSigLIP** falls further to mAP **0.9226** and P@1 **0.9335**. The SigLIP–MedSigLIP gap mirrors Lung Histo, suggesting medical adaptation does not consistently help colon retrieval and may reduce embedding discriminability.

A clear performance decline is observed for segmentation-oriented encoders. **SAM ViT-B** attains mAP **0.9706** and P@1 **0.9840**, indicating that segmentation-optimized features are less effective for fine-grained histopathology retrieval than contrastive or discriminative pretraining. The most pronounced degradation is seen for **MedSAM ViT-B**, which falls far behind all other backbones with mAP **0.6301**, MRR **0.7059**, and P@1 **0.5230**. This gap confirms that segmentation-focused adaptation alone does not ensure strong retrieval embeddings, as mask-driven features may not preserve global similarity for nearest-neighbor ranking.

Overall, the Colon Histo results confirm that **UNI** is uniquely well-suited to histopathology retrieval, achieving the only perfect benchmark score across all experiments. The near-perfect performance of **Swin-B**, **CLIP ViT-L/14**, **DINOv2 ViT-B/14**, **BiomedCLIP**, and **BMC-CLIP-CF** further demonstrates that large-scale pretrained semantic embeddings transfer extremely well to colon histopathology patch retrieval. In contrast, sigmoid-based contrastive models and segmentation-oriented encoders show reduced effectiveness, with **MedSAM ViT-B** performing particularly poorly. These findings show that histopathology retrieval depends less on specialization and more on embeddings that preserve discriminative morphological similarity.

4.4 Wireless Capsule Endoscopy Retrieval

The WCE dataset is different from the histopathology dataset, being based on endoscopic images. Thus, the retrieval errors in this setting appear from non-diagnostic variability (illumi-

nation shifts, different viewpoints, and mucosal structure). As a result, this dataset provides a useful test of cross-domain robustness and sensitivity to acquisition-level variation.

On WCE, the best overall performance is achieved by **BMC-CLIP-CF**, which reaches an mAP of **0.9809**, MRR of **0.9927**, P@1 of **0.9892**, P@5 of **0.9697**, and P@10 of **0.9620**. The second-best result is obtained by **OpenCLIP ViT-H/14** (mAP **0.9698**, P@1 **0.9775**), which, despite being a general-purpose encoder trained on open-source data, places ahead of all domain-specific models except BMC-CLIP-CF. This suggests that scale and capacity within the contrastive pretraining family can compensate for the absence of biomedical domain adaptation in endoscopic retrieval. **Swin-B** is the strongest standard backbone (mAP **0.9684**, P@1 **0.9850**), followed closely by **BiomedCLIP** (mAP **0.9668**) and **CLIP ViT-L/14** (mAP **0.9644**). **ResNet50** and **DINOv2 ViT-B/14** form a competitive middle tier (mAP **0.9573** and **0.9565**, respectively), while **ViT-B/16** performs slightly lower at **0.9486** mAP. **UNI** reaches **0.9521** mAP, a respectable result for a pathology-specific model evaluated outside its pretraining domain, although it does not match the top contrastive encoders.

A notable finding concerns SigLIP and MedSigLIP. Despite sharing the sigmoid contrastive pretraining objective, their performance on WCE is substantially lower than that of CLIP-family models. **SigLIP** achieves **0.9203** mAP and **MedSigLIP** only **0.8935** mAP, placing them among the weaker models on this dataset. This is somewhat unexpected given SigLIP's strong general-purpose performance reported elsewhere, and may reflect sensitivity to the image resolution or preprocessing differences inherent to endoscopic imagery. The further drop from SigLIP to MedSigLIP suggests that the medical adaptation in this case does not improve, and may slightly harm, transferability to gastrointestinal endoscopy.

A clear performance drop is observed for reconstruction and segmentation-oriented encoders. **MAE ViT-B/16** reaches **0.9131** mAP and **SAM ViT-B** attains **0.9207** mAP, indicating that features optimized for masked reconstruction or promptable segmentation do not transfer as effectively to retrieval in unconstrained endoscopic imagery. The most pronounced degradation is seen for **MedSAM ViT-B**, which falls far behind all other models with **0.5512** mAP and **0.4325** P@1. This large gap suggests that specialization toward medical segmentation alone is insufficient for instance retrieval when appearance variability is driven by acquisition artifacts and viewpoint changes rather than stable structural cues.

Overall, the WCE results confirm that representation quality is domain-sensitive, and that the backbone ranking does not strictly mirror what is observed in histopathology. Strong contrastive models, particularly BMC-CLIP-CF and OpenCLIP ViT-H/14, transfer well to endoscopic data, while sigmoid-based contrastive models and domain-specialized encoders for segmentation show reduced effectiveness. These findings indicate that success in microscopy does not automatically imply equal performance on endoscopic imagery, as WCE retrieval places greater emphasis on robustness to acquisition conditions and scene-level variability rather than solely on fine-grained tissue morphology.

4.5 Merged Dataset Retrieval

The merged benchmark combines Lung Histopathology, Colon Histopathology, and Wireless Capsule Endoscopy (WCE) datasets together. Their different visual characteristics makes the task harder. Consequently, intra-domain fine-grained variability (tissue morphology) and inter-domain shifts (microscopy and endoscopic data) are introduced. In this case, retrieval errors can appear from a mixture of semantic confusion, domain imbalance, and dataset acquisition. As a result, the merged dataset provides a comprehensive test of both representation generalization and cross-domain robustness.

The results are presented in TABLE 2. Across all models, the best overall performance is achieved by **BMC-CLIP-CF**, which outperforms all other backbones on every metric, reaching an mAP of **0.9947**, MRR of **0.9979**, P@1 of **0.9969**, P@5 of **0.9926**, and P@10 of **0.9884**. This result showcases the strong capacity of contrastive training in capturing fine-grained features and broader visual patterns.

The second-best performance is observed for **CLIP ViT-L/14**, which achieves an mAP of **0.9928**, MRR of **0.9975**, and P@1 of **0.9961**, closely followed by **Swin-B** (mAP **0.9920**) and **BiomedCLIP** (mAP **0.9921**). These models form a strong baseline, indicating that both large-scale contrastive pretraining and hierarchical transformer architectures are effective in handling mixed-domain retrieval. **OpenCLIP ViT-H/14** also performs competitively (mAP **0.9902**, MRR **0.9972**), demonstrating that increased model capacity can compensate for the lack of domain-specific adaptation even in a heterogeneous setting.

In the mixed dataset setting, a moderate-performing group, including **UNI**, **DINOv2 ViT-B/14**, and **ResNet50**, demonstrates competitive retrieval performance across heterogeneous domains. This suggests that both self-supervised approaches (**DINOv2**) and traditional convolutional networks (**ResNet50**) retain strong generalization despite domain variability. In contrast, **ViT-B/16** performs slightly lower, indicating that smaller transformer models may struggle to fully capture the diversity present in the merged dataset.

A noticeable drop is observed for sigmoid-based contrastive models and reconstruction-based approaches. **SigLIP** achieves an mAP of **0.9768**, while **MedSigLIP** further declines to **0.9376**, suggesting that the sigmoid objective may be less robust in highly heterogeneous retrieval settings. Similarly, **MAE ViT-B/16** reaches **0.9750** mAP, indicating that representations optimized for masked reconstruction do not transfer as effectively to retrieval tasks involving multiple visual domains.

The weakest performance is obtained by the segmentation-based models. **SAM ViT-B** achieves **0.9600** mAP, while **MedSAM ViT-B** exhibits a substantial downgraded performance, with only **0.5377** mAP and **0.4260** P@1, compared to the contrastive models. This large performance drop suggests that features learned for promptable segmentation, particularly when specialized for medical tasks, do not generalize well to instance retrieval across diverse imaging modalities.

Overall, the merged dataset results reinforce the importance of both scale and training objective. Contrastive models, especially BMC-CLIP-CF and CLIP-based variants, demonstrate strong cross-domain generalization, while reconstruction- and segmentation-focused

encoders show limited transferability. The findings further indicate that success in individual domains does not directly translate to optimal performance in heterogeneous settings, where robustness to both semantic and acquisition-level variation becomes critical.

4.6 Qualitative Retrieval Examples

FIGURE 2 – FIGURE 9, present representative query images together with their top retrieved neighbors across the Lung Histo, Colon Histo, WCE datasets, and merged datasets. For each setting, one figure shows a PCA projection of the embeddings colored by cosine similarity to the query, highlighting visually similar images around the query, while the other figure shows embeddings colored by class labels, emphasizing semantic separation between tissue types. The top retrieved images are highlighted with larger markers, and the query is marked with a red star in all plots. These qualitative examples complement the quantitative retrieval metrics by illustrating how well the embedding spaces capture visual and morphological similarity, allowing assessment of whether retrieval errors are arbitrary or remain visually and semantically meaningful.

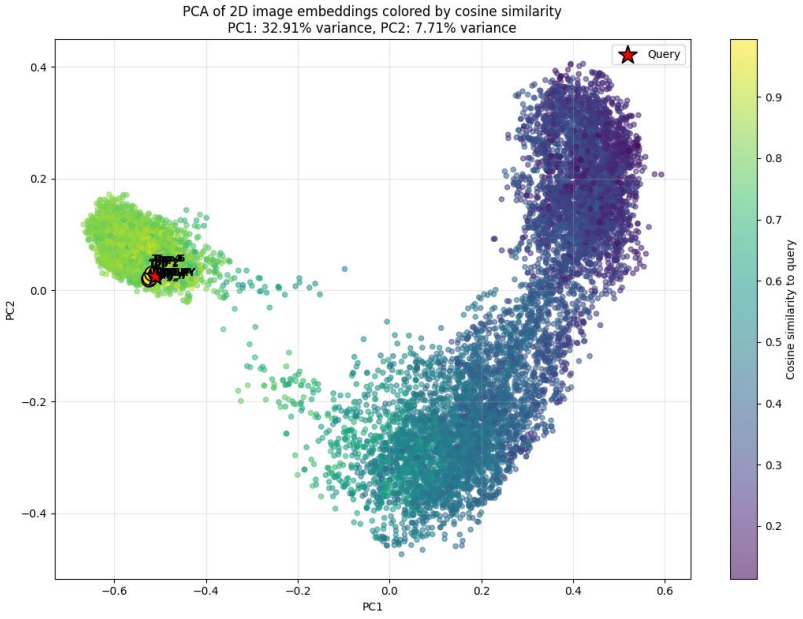


Figure 2: Lung Histo - using the UNI backbone – Cosine Similarity PCA. PCA of Lung Histo image embeddings colored by cosine similarity to a selected query. The top retrieved images are highlighted with larger markers, and the query image is marked with a red star.

The qualitative analysis should discuss whether incorrect retrievals still preserve meaningful visual similarity. In medical image retrieval, a false positive may remain informative if it reflects related morphology, neighboring pathological patterns, or clinically plausible appearance overlap rather than a completely unrelated mismatch.

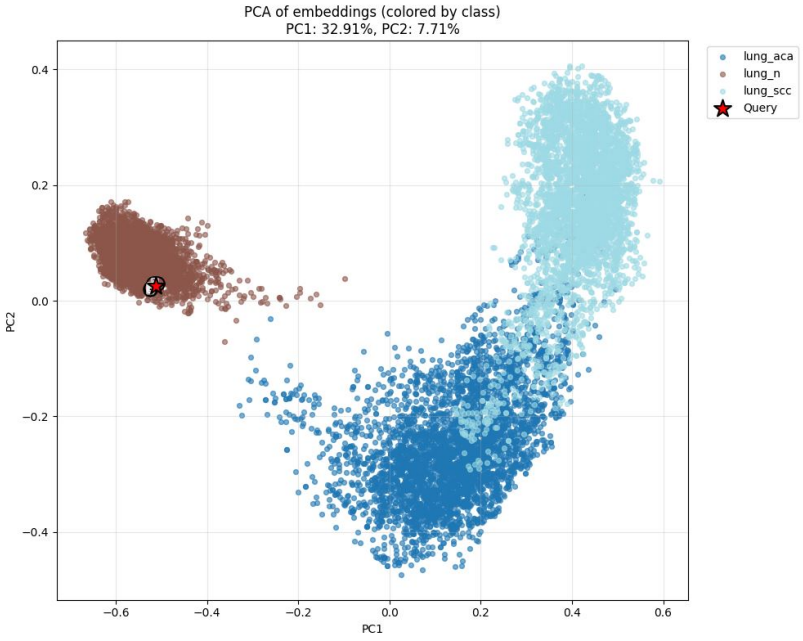


Figure 3: Lung Histo – using the UNI backbone - Class-colored PCA. PCA of Lung Histo embeddings colored according to class labels. The top retrieved images are highlighted, and the query is marked with a red star.

4.7 Extended Evaluation on the WCE Dataset

4.7.1 Setup

We extend our experimental evaluation on the **WCE dataset** with additional statistical analysis, fine-tuning experiments, and systematic failure analysis.

First, we incorporate **statistical uncertainty estimation** by reporting **95% confidence intervals** for all retrieval metrics using bootstrap resampling over queries. In addition, we perform **paired statistical tests** (bootstrap and randomization tests) between models, ensuring fair comparison since all methods are evaluated on the same query set.

Second, we conduct a comprehensive set of **fine-tuning experiments on WCE** across multiple backbones, including ResNet50, ViT-B/16, Swin-B, CLIP ViT-L/14, SigLIP, and OpenCLIP ViT-H/14. We evaluate several adaptation strategies: linear probing, metric-aware fine-tuning, and full end-to-end fine-tuning.

Third, we introduce a **systematic failure analysis** on WCE, including per-class performance breakdown, confusion pattern analysis, hardest-query identification, and embedding space analysis (intra-class and inter-class similarity).

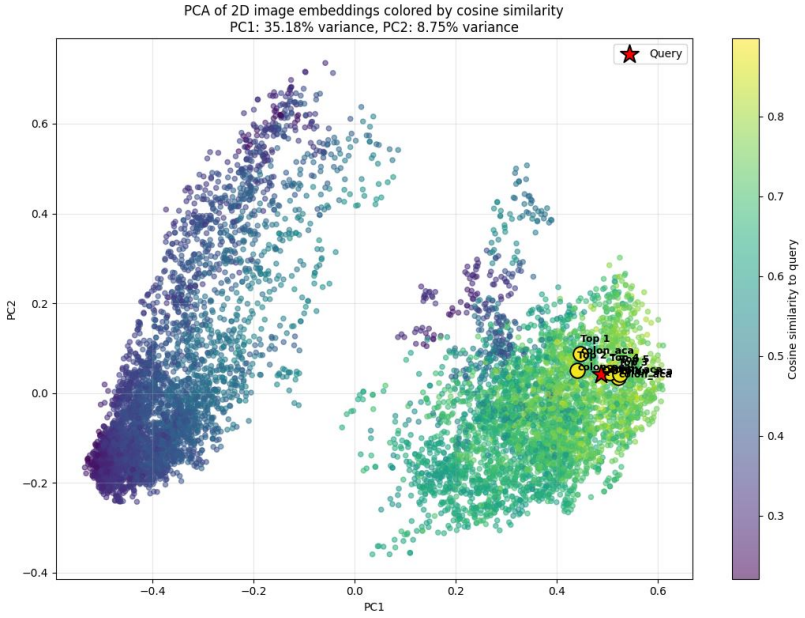


Figure 4: Colon Histo - using the UNI backbone – Cosine Similarity PCA. PCA of Colon Histo image embeddings colored by cosine similarity to a selected query. The top retrieved images are highlighted with larger markers, and the query image is marked with a red star.

4.7.2 Results

On the **WCE dataset**, fine-tuning plays a critical role in retrieval performance. While frozen features already provide strong baselines, all architectures benefit significantly from supervised adaptation. For example, **ResNet50** improves from an mAP of 0.955 to 0.982 after fine-tuning, while **ViT-B/16** increases from 0.940 to 0.993. Similar trends are observed across all transformer-based models. In FIGURE 10, we visualize the PCA projection of retrieval embeddings for ResNet50 using frozen features versus full fine-tuning, highlighting how fine-tuning changes the feature space structure and nearest-neighbor retrieval behavior.

The best overall performance on WCE is achieved by **CLIP ViT-L/14 with full fine-tuning**, which reaches an mAP of 0.9966, MRR of 0.9969, and P@1 of 0.9958. Close competitors include **Swin-B with metric fine-tuning** (mAP 0.9949) and **ViT-B/16 with full fine-tuning** (mAP 0.9934), indicating that both hierarchical transformers and standard ViT architectures can achieve near-perfect retrieval when properly adapted.

A key observation from the WCE experiments is that **metric-aware fine-tuning provides performance comparable to full end-to-end training**, while requiring fewer trainable parameters. This suggests that explicitly optimizing the embedding space is highly effective for retrieval tasks.

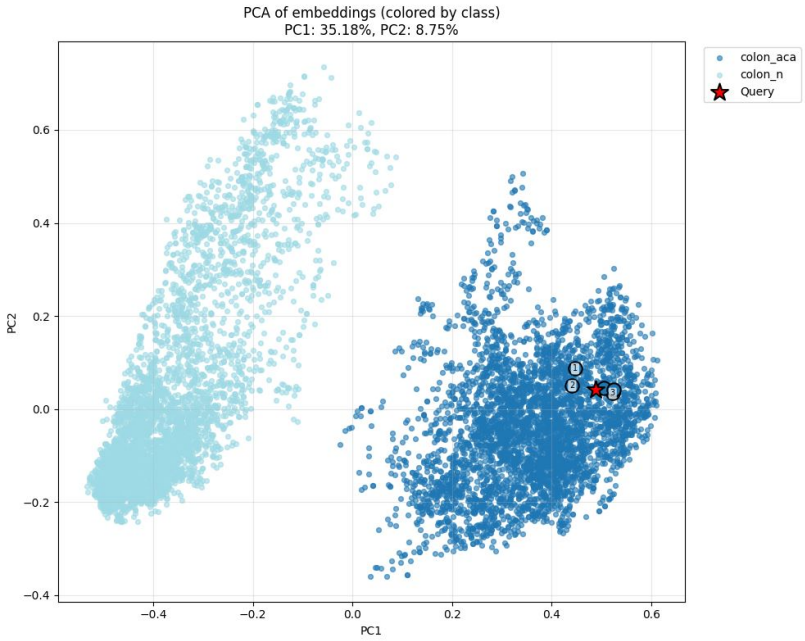


Figure 5: Colon Histo – using the UNI backbone - Class-colored PCA. PCA of Colon Histo embeddings colored according to class labels. Top retrieved images are highlighted, and the query is marked with a red star.

Despite their domain-specific design, biomedical models such as BiomedCLIP do not consistently outperform large general-purpose contrastive models on WCE. Instead, **model capacity and pretraining scale appear more important than domain specialization**, particularly when combined with fine-tuning.

In contrast, reconstruction- and segmentation-based encoders exhibit significantly weaker performance. MAE and SAM achieve mAP values around $0.91 - 0.92$, while MedSAM drops drastically to 0.56. This confirms that representations optimized for reconstruction or segmentation do not transfer effectively to retrieval on WCE.

To ensure the robustness of our conclusions on the WCE dataset, we complement the quantitative results with statistical and systematic analyses. We report **95% confidence intervals** for all retrieval metrics using bootstrap resampling over the 1200 query images, and perform **paired statistical tests** (bootstrap and randomization) between models.

Although fine-tuning improves the point estimates across several backbones, the paired statistical tests do not indicate statistically significant differences at the 0.05 level with respect to the best-performing model (CLIP ViT-L/14 with full fine-tuning). This suggests that the benchmark operates in a saturated regime, where performance differences are small relative to the variability across queries.

In addition, we conduct a **systematic failure analysis**, including per-class performance, confusion patterns, and hardest-query identification. The results show that the remaining errors

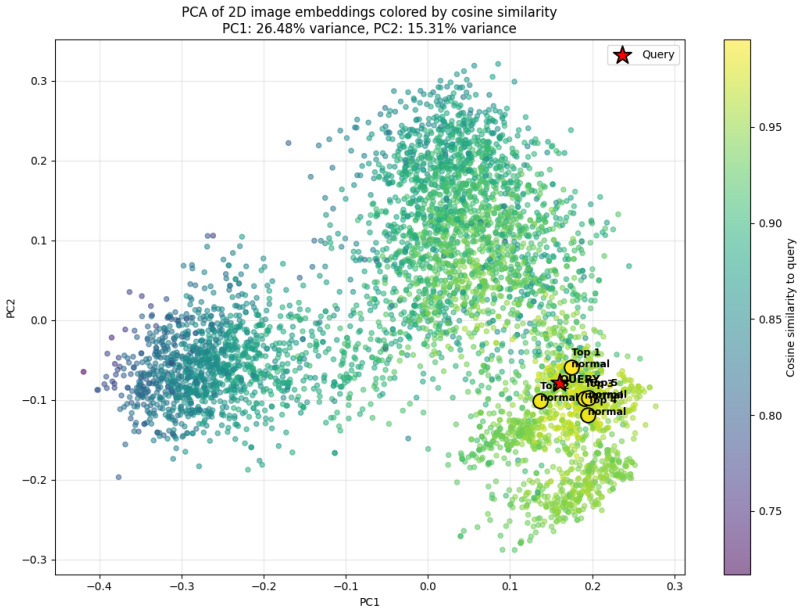


Figure 6: WCE – using the BMC-CLIP-CF - Cosine Similarity PCA. PCA of WCE image embeddings colored by cosine similarity to a selected query. Top retrieved images are highlighted with larger markers, and the query is marked with a red star.

are concentrated in a small subset of difficult queries, typically corresponding to visually ambiguous cases with high inter-class similarity. Furthermore, embedding similarity analysis reveals a strong separation between intra-class and inter-class distributions, explaining the overall high retrieval performance while highlighting the difficulty of the remaining failure cases.

4.8 Test Settings and Learning Strategy

All experiments were conducted on a 2D image retrieval benchmark constructed from a class-organized dataset, where images were stored in class-specific folders. The dataset was divided using a stratified 80/20 train–test split in order to preserve the class distribution across both subsets. For the computational time analysis, a stratified 5% subset of the dataset was additionally considered to benchmark runtime and estimate the scaling behavior for the full 100% dataset. During evaluation, embeddings were extracted for both the training and test images, and each test image was used as a query against the training set gallery. Retrieval was performed in the normalized embedding space by selecting the top-K most similar training samples, with $K \in \{1, 5, 10\}$. Performance was measured using Precision@K, HitRate@K, Recall@K, mean Average Precision (mAP), and Mean Reciprocal Rank (MRR).

The evaluated backbones included both conventional and foundation-based visual encoders, namely ResNet50, ViT-B/16, Swin-B, CLIP ViT-L/14, BiomedCLIP, DINOv2 ViT-B/14, MAE ViT-B/16, BMC-CLIP-CF, SAM ViT-B, MedSAM ViT-B, SigLIP, MedSigLIP, UNI,

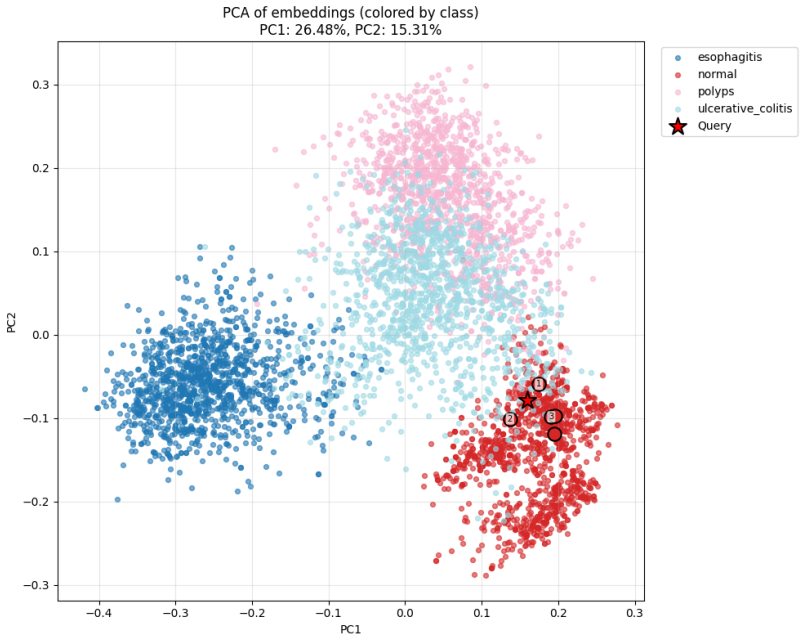


Figure 7: WCE – using the BMC-CLIP-CF - Class-colored PCA. PCA of WCE embeddings colored by class labels. Top retrieved images are highlighted, and the query is marked with a red star.

and OpenCLIP ViT-H/14. Input images were initially resized to 224×224, while each backbone internally adapted the input to its preferred resolution when required. The batch size was set to 32, the number of data loading workers to 2, and the final embedding dimensionality to 512. Inference in bfloat16 precision was enabled for compatible models in order to improve computational efficiency. To ensure reproducibility, all experiments were performed with a fixed random seed of 42.

Regarding the learning strategy, no task-specific training or fine-tuning was performed. All backbones were used as frozen pre-trained feature extractors, meaning that their parameters remained unchanged throughout the evaluation process. When the native feature dimensionality of a model differed from the target embedding size, a linear projection layer was used to map the extracted features to a common 512-dimensional representation space. Finally, all embeddings were L2-normalized before similarity computation. Therefore, the adopted strategy can be characterized as a training-free transfer evaluation protocol designed to assess the intrinsic retrieval capability of different pre-trained visual encoders.

4.9 Computational Time Analysis

To better understand the computational cost of the evaluated retrieval backbones, we analyzed the runtime in terms of *average processing time per image*. The experiments were conducted on a CUDA-enabled GPU. Since embedding extraction dominates the overall

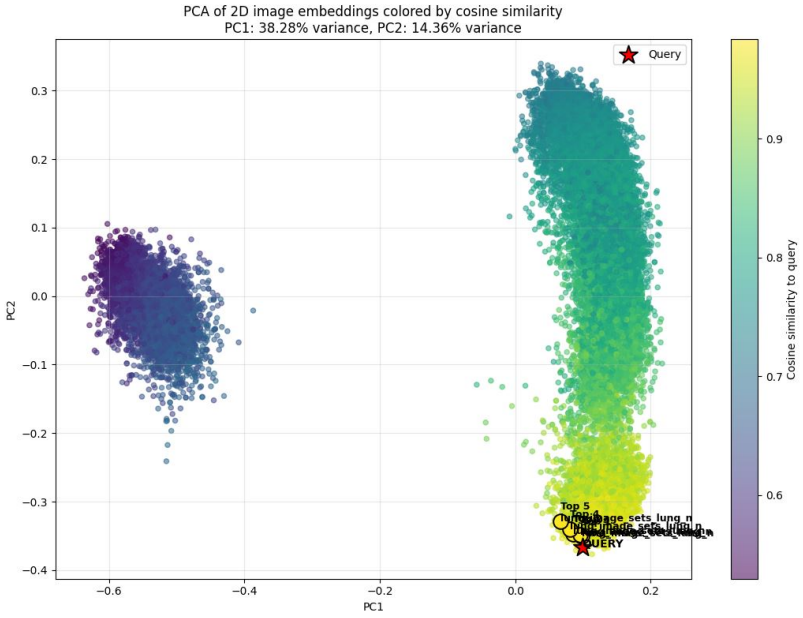


Figure 8: Merged datasets – using the BMC-CLIP-CF - Cosine Similarity PCA. PCA of merged datasets' image embeddings colored by cosine similarity to a selected query. Top retrieved images are highlighted with larger markers, and the query is marked with a red star.

inference pipeline, the per-image analysis provides a clearer comparison of the practical efficiency of each backbone.

The results show that lightweight classification-oriented backbones were the most efficient. In particular, ViT-B/16, Swin-B, SigLIP, ResNet50, MAE ViT-B/16, and BiomedCLIP are suitable for fast large-scale feature extraction. Among these models, ViT-B/16 and SigLIP were the fastest, with an average embedding time of roughly $0.022 - 0.024$ seconds per image, while ResNet50 and BiomedCLIP remained similarly efficient at around $0.026-0.028$ seconds per image. Slightly larger transformer models such as UNI and CLIP ViT-L/14 were slower, requiring approximately 0.059 and 0.076 seconds per image, respectively. OpenCLIP ViT-H/14 and BMC-CLIP-CF were substantially heavier, reaching about 0.139 and 0.088 seconds per image.

A markedly different behavior was observed for the segmentation-based or very large medical foundation models. SAM ViT-B and MedSAM ViT-B required approximately 0.458 and 0.515 seconds per image, respectively, while MedSigLIP required around 0.450 seconds per image. These values indicate that such models are roughly one order of magnitude slower than the more efficient classification-oriented encoders. Therefore, although these models may offer richer visual representations, their computational cost is significantly higher when used as generic image embedding extractors in a retrieval setting.

From a practical point of view, this highlights an important trade-off between representational strength and computational efficiency. Smaller and medium-scale backbones provide

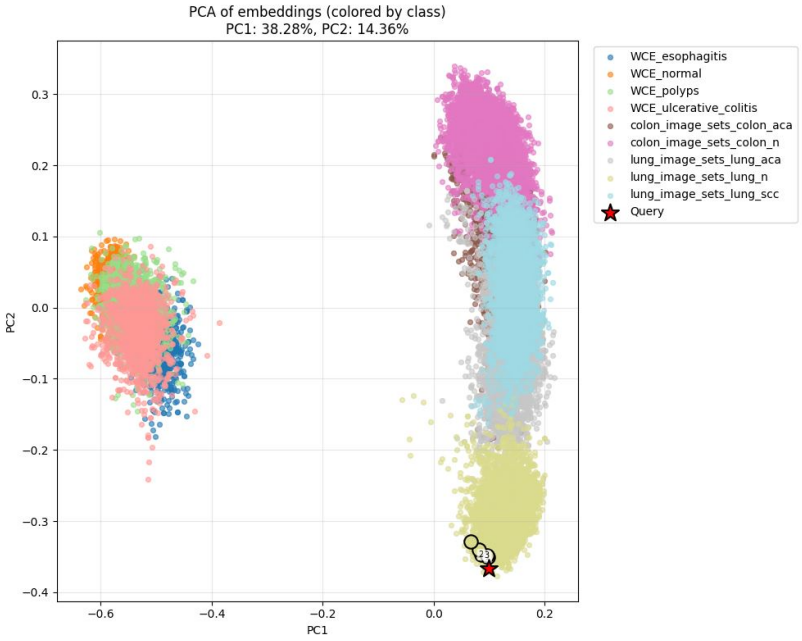
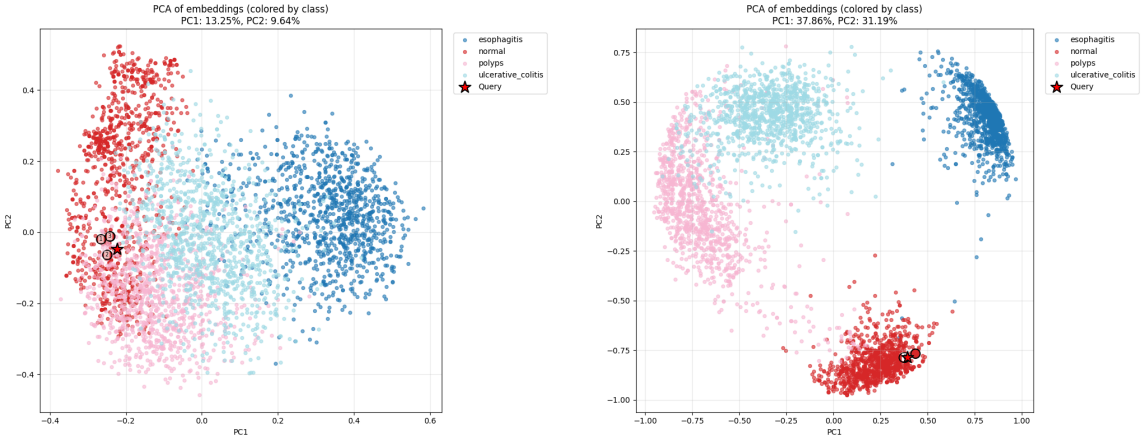


Figure 9: Merged datasets – using the BMC-CLIP-CF - Class-colored PCA. PCA of merged datasets' embeddings colored by class labels. Top retrieved images are highlighted, and the query is marked with a red star.



Frozen ResNet50 Fine-tuned ResNet50

Figure 10: Comparison of embeddings between frozen and fine-tuned ResNet50.

much faster per-image processing and are therefore more appropriate when rapid embedding generation or large-scale deployment is required. In contrast, heavier architectures such as SAM, MedSAM, MedSigLIP, and OpenCLIP ViT-H/14 have a substantially higher cost per image, which may limit their usability in time-sensitive or resource-constrained applications.

Consequently, the choice of backbone should not depend only on retrieval effectiveness, but also on the acceptable inference time budget per image.

5. DISCUSSION

This study benchmarks a diverse set of pretrained backbones for CBIR across three medical image domains that differ substantially in visual structure, acquisition conditions, and semantic variability. By evaluating all models as frozen feature extractors, the analysis isolates representation quality and makes the results directly relevant to low-label settings, fast deployment, and retrieval-based clinical workflows.

A consistent pattern across the experiments is the strong domain dependence of retrieval performance. The two histopathology benchmarks, Lung Histo and Colon Histo, are both characterized by highly structured visual patterns driven by tissue morphology, staining, and cellular organization. In these settings, retrieval is largely determined by fine-grained texture and micro-architectural cues rather than viewpoint or acquisition variability.

On Lung Histo, performance is already close to saturation for most modern backbones. The best result is achieved by **UNI**, which reaches near-perfect scores across all metrics, reflecting the advantage of large-scale, domain-specific self-supervised pretraining when the target task closely matches the training distribution. Close behind are **CLIP ViT-L/14** and **Swin-B**, followed by **BMC-CLIP-CF**, **DINOv2 ViT-B/14**, **ResNet50**, and **BiomedCLIP**, all of which achieve nearly indistinguishable performance. This tight clustering indicates that many large-scale pretrained encoders are already sufficient to capture discriminative morphological similarity in lung histopathology.

A similar but even more extreme trend is observed on Colon Histo. Here, **UNI** achieves perfect retrieval performance across all metrics, the only such result in the benchmark. Several other models, including **Swin-B**, **CLIP ViT-L/14**, **DINOv2 ViT-B/14**, **BiomedCLIP**, and **BMC-CLIP-CF**, also reach effectively saturated performance with near-perfect ranking. Even **ResNet50** remains highly competitive, suggesting that this dataset is strongly separable at the feature level. As a result, Colon Histo offers limited resolution for distinguishing between top-performing backbones, since most models already produce embeddings that are sufficient for near-perfect nearest-neighbor retrieval.

In contrast, the WCE benchmark presents a different and more challenging scenario. Unlike histopathology, the endoscopic imagery is affected by several changes (illumination, blur, viewpoint, and visual content). Thus, the retrieval depends less on stable micro-structure and more on robustness to acquisition variability and global semantic consistency.

As per our experiments, the best results on WCE are achieved by **BMC-CLIP-CF**, followed by **OpenCLIP ViT-H/14** and **Swin-B**, with **BiomedCLIP** and **CLIP ViT-L/14** close behind. Notably, **OpenCLIP ViT-H/14**, a general-purpose large-scale contrastive model, outperforms most domain-specific encoders, suggesting that model scale and representation capacity can compensate for the lack of medical specialization in this domain. Mid-tier performance is observed for **ResNet50**, **DINOv2 ViT-B/14**, and **UNI**, the latter showing

reasonable transfer despite being trained on histopathology data. Overall, the WCE dataset introduces a clearer separation between models. Also, it highlights differences that are largely hidden in the histopathology benchmarks.

Across all three datasets, the pretraining strategy plays a central role in determining retrieval quality. Contrastive and multimodal models, including CLIP variants, **BiomedCLIP**, and **BMC-CLIP-CF**, consistently rank among the top performers, indicating that alignment between visual embeddings and semantic similarity is critical for nearest-neighbor retrieval. Strong hierarchical transformers such as **Swin-B** also perform robustly across domains, while self-supervised models like **DINOv2** remain highly competitive, especially in morphology-driven tasks, confirming that explicit label supervision is not strictly required for high-quality representations.

The reconstruction-based and segmentation-oriented encoders show weaker transfer to retrieval. **MAE ViT-B/16** remains competitive but consistently below the top group, particularly on WCE, suggesting that masked reconstruction objectives do not fully preserve discriminative ranking structure. The gap is larger for **SAM ViT-B**, and most pronounced for **MedSAM ViT-B**, which performs significantly worse across all datasets. This consistent degradation indicates that features optimized for mask prediction do not naturally translate into compact, globally discriminative embeddings suitable for similarity search.

Sigmoid-based contrastive models (**SigLIP** and **MedSigLIP**) exhibit mixed behavior. While they perform reasonably well on histopathology, they consistently fall behind standard CLIP-style models and degrade more noticeably on WCE. The medical adaptation in **MedSigLIP** does not provide a clear advantage and, in some cases, reduces performance, suggesting that not all forms of domain-specific modification improve retrieval quality.

Several practical considerations should be taken into account when interpreting these results. First, the histopathology datasets, particularly Colon Histo, exhibit near-saturated performance, which limits their ability to differentiate between strong backbones. Second, both histopathology benchmarks rely on patch-based data that may include correlations introduced during dataset construction or augmentation. Third, the evaluation uses image-level splits rather than patient- or slide-level separation, due to dataset constraints. Finally, all models are evaluated without fine-tuning, which ensures comparability but may underestimate the potential of methods that benefit from domain adaptation.

Overall, the benchmark highlights three key insights. First, retrieval performance is strongly domain-dependent, and success in histopathology does not guarantee optimal performance in endoscopic imaging. Second, large-scale contrastive and multimodal pretraining produces the most robust and transferable embeddings across diverse medical domains. Third, specialization toward segmentation or reconstruction alone is insufficient for CBIR, as effective retrieval requires embeddings that preserve global semantic similarity for ranking. These findings establish a clear baseline for future work on fine-tuning, hybrid representations, and retrieval-augmented clinical decision systems.

6. CONCLUSIONS

In the present work, we propose a unified pipeline for content-based image retrieval across two histopathology datasets and one wireless capsule endoscopy dataset. Our benchmark evaluates different pretrained backbones under the same experimental protocol (frozen versus unfrozen feature retrieval). To evaluate the performance of the retrieval methods, we report standard metrics, including mean Average Precision (mAP), Mean Reciprocal Rank (MRR), and Precision@K.

Our framework is designed to address a practical and clinically relevant question: which backbone provides the most effective image embeddings for medical image retrieval across diverse domains? We are treating lung histopathology, colon histopathology, and wireless capsule endoscopy (WCE) as distinct retrieval tasks, and in parallel, we combine the datasets together. The present study enables a granular analysis of cross-domain robustness, representation quality, and model transferability. Our results present how different imaging modalities impose different demands on feature representations, ranging from fine-grained morphological discrimination in histopathology to robustness against acquisition variability in endoscopy.

Beyond benchmarking, the current framework establishes a foundation for several research directions. Hybrid approaches that combine complementary backbones or fuse multi-scale features may help bridge the gap between texture-sensitive and semantics-driven representations. Another direction, re-ranking techniques, and retrieval-augmented pipelines could be integrated to improve downstream clinical decision support. Finally, expanding the benchmark to include patient-level splits, additional modalities, and multimodal retrieval (image-text alignment) would provide a more realistic assessment of model performance in real-world clinical settings.

7. DATA AVAILABILITY STATEMENT

The datasets used in this study are publicly available from their respective original sources. Additional implementation details and derived outputs are available from the corresponding author upon reasonable request. The Python implementation of the method described in this paper is available at GitHub.

8. ACKNOWLEDGMENTS

The author acknowledges the developers of the public datasets and open-source models used in this study.

References

[1] Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. Comput Struct Biotechnol J. 2018;16:34-42.

[2] Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, et al. A Survey on Deep Learning in Medical Image Analysis. Med Image Anal. 2017;42:60-88.

[3] Zhu S, Phan J, Chennupati S, Duong D, Nguyen T, Dinh A, et al. Public Imaging Datasets of Gastrointestinal Endoscopy on Artificial Intelligence: A Systematic Review. Diagnostics. 2023;13:3136.

[4] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE. 2016:770-778.

[5] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, et al. An Image Content-Based Image Retrieval in Histopathology and Gastrointestinal Endoscopy Is Worth. In: Proceedings of the international conference on learning representations (ICLR), virtual event. Vol. 16×16 words: Transformers for image recognition at scale. 2021.

[6] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), Montreal, Canada. IEEE/CVF. 2021:10012-10022.

[7] Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. DINOv2: Learning Robust Visual Features Without Supervision. 2023. ArXiv Preprint: https://arxiv.org/pdf/2304.07193

[8] He K, Chen X, Xie S, Li Y, Dollár P, et al. Masked Autoencoders Are Scalable Vision Learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE/CVF. 2022:16000-16009.

[9] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th international conference on machine learning (ICML 2021), virtual event. ICML. 2021:8748-8763.

[10] Zhai X, Mustafa B, Kolesnikov A, Beyer L. Sigmoid Loss for Language Image Pretraining. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV). IEEE/CVF. 2023:11975-11986.

[11] https://zenodo.org/records/5143773

[12] Zhang S, Xu Y, Usuyama N, Xu H, Bagga J, Tinn R, et al. BiomedCLIP: A Multimodal Biomedical Foundation Model Pretrained From Fifteen Million Scientific Image-Text Pairs. 2023. ArXiv Preprint: https://arxiv.org/pdf/2303.00915

[13] Chen RJ, Ding T, Lu MY, Williamson DF, Jaume G, Song AH, et al. Towards a General-Purpose Foundation Model for Computational Pathology. Nat Med. 2024;30:850-862.

[14] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment Anything. 2023. ArXiv Preprint: https://arxiv.org/pdf/2304.02643

[15] Ma J, He Y, Li F, Han L, You C, Wang B. Segment Anything in Medical Images. Nat Commun. 2024;15:654.

[16] Mahbod A, Saeidi N, Hatamikia S, Woitek R. Evaluating Pre-Trained Convolutional Neural Networks and Foundation Models as Feature Extractors for Content-Based Medical Image Retrieval. Eng Appl Artif Intell. 2025;150:110571.

[17] Nanni L, Lumini A, Brahnam S. Local Binary Patterns Variants as Texture Descriptors for Medical Image Analysis. Artif Intell Med. 2010;49:117-125.

[18] Ergen B, Baykara M. Texture Based Feature Extraction Methods for Content Based Medical Image Retrieval Systems. Biomed Mater Eng. 2014;24:3055-3062.

[19] Kumar A, Kim J, Cai W, Fulham M, Feng D. Content-Based Medical Image Retrieval: A Survey of Applications to Multidimensional and Multimodality Data. J Digit Imaging. 2013;26:1025-1039.

[20] Sellergren A, Kazemzadeh S, Jaroensri T, Kiraly A, Traverse M, Kohlberger T, et al.MedGemma Technical Report. 2025. ArXiv preprint: https://arxiv.org/pdf/2507.05201

[21] Borkowski AA, Bui MM, Thomas LB, Wilson CP, DeLand LA, Mastorides SM. Lung and Colon Cancer Histopathological Image Dataset (LC25000). 2019. ArXiv Preprint: https://arxiv.org/pdf/1912.12142

[22] Lozano A, Tabassum A, Wang J, Hu M, Liang Y, Bissonnette V, et al. Biomedica: An Open Biomedical Image-Caption Archive [Dataset], and Vision-Language Models Derived From Scientific Literature. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE/CVF. 2025.