



OPEN ACCESS

EDITED BY

Palak Handa,
Danube Private University, Austria

REVIEWED BY

Kai Zhang,
Chongqing Chang'an Industrial Co., Ltd.,
China
Imran Iqbal,
Helmholtz Association of German Research
Centres (HZ), Germany
Nima Torbati,
Danube Private University, Austria

*CORRESPONDENCE

Youngbae Hwang
✉ ybhwang@cbnu.ac.kr

[†]These authors have contributed equally to
this work and share first authorship

RECEIVED 18 October 2025

REVISED 19 November 2025

ACCEPTED 26 November 2025

PUBLISHED 18 December 2025

CITATION

Kim K, Park J, Kim SH and Hwang Y (2025)
Improving image-retrieval performance of
foundation models in gastrointestinal
endoscopic images.
Front. Med. 12:1727884.
doi: 10.3389/fmed.2025.1727884

COPYRIGHT

© 2025 Kim, Park, Kim and Hwang. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

Improving image-retrieval performance of foundation models in gastrointestinal endoscopic images

Kangsan Kim^{1†}, Junseok Park^{2†}, Sang Hyun Kim³ and Youngbae Hwang^{1*}

¹Department of Intelligent Systems and Robotics, Chungbuk National University, Cheongju, Republic of Korea, ²Department of Internal Medicine, Soonchunhyang University College of Medicine, Seoul, Republic of Korea, ³Department of Surgery, Soonchunhyang University College of Medicine, Seoul, Republic of Korea

Introduction: The quality of gastrointestinal endoscopy is verified by documenting specific required images, but identifying these images from the numerous photographs captured during a procedure is tedious. Conventional deep-learning approaches that aim to automate this process are often limited by subjective assessments and poor interpretability.

Methods: We introduce a novel content-based image-retrieval framework that employs a dual-backbone architecture, integrating a general-purpose vision foundation model (DINOv2) and a domain-specific endoscopic model (GastroNet). The system is trained using parameter-efficient metric learning to generate discriminative embeddings for efficient similarity searches. The framework is evaluated on 3,500 public endoscopic images (from the Kvasir and HyperKvasir datasets) and validated on entirely unseen real-world and synthetic data.

Results: Our model achieves state-of-the-art performance (97.71% Recall@1, 99.14% Recall@5, and 96.74% mean average precision), which is significantly superior to those of single-backbone baseline models. Ablation studies confirm that this improvement is primarily due to the two backbones capturing complementary features.

Discussion: These findings demonstrate that the proposed dual-backbone framework offers an accurate and automated tool for assessing the procedural quality of gastrointestinal endoscopy and may facilitate more reliable quality control in clinical practice.

KEYWORDS

gastrointestinal endoscope, artificial intelligence, deep learning, image retrieval, foundation model

1 Introduction

Endoscopy is a critical diagnostic tool for gastrointestinal diseases, which enables the direct visual inspection of internal organs. For instance, during esophagogastroduodenoscopy (EGD), the entire stomach cannot be visualized in a single field of view; therefore, images of specific anatomical segments must be captured (1). These landmark images serve as crucial quality control indicators (2, 3). Currently, manually verifying these images from the numerous photographs taken is a tedious task, prompting the automation of the process (4, 5). However, traditional deep-learning-based approaches are hindered by the need for large, comprehensively annotated training datasets, and their susceptibility to overfitting limits their performance across diverse data (6–8).

As an alternative, content-based image retrieval (CBIR) offers a more flexible and interpretable paradigm for analyzing endoscopic images (9, 10). Instead of relying on predefined labels, CBIR operates by comparing the extracted features of a query image with those in a reference database to find similar items (11, 12). This approach enhances clinical explainability and better accommodates image variability; its performance can be further improved using supplementary techniques, such as triplet loss (13). Recent advancements in CBIR involve utilizing large-scale foundation models, which offer advantages such as reduced training requirements and improved generalizability (14).

Accordingly, this study proposes a framework that leverages a foundation model for endoscopic image retrieval. This system enables efficient search for similar images within endoscopic image datasets, guided by anatomically relevant features and contextual similarity to a user-defined reference image.

2 Materials and methods

The proposed CBIR system for gastrointestinal endoscopy images (Figure 1) processes a given query image by mapping it into a low-dimensional feature vector or embedding. It then retrieves the most visually similar images from a database by ranking their embeddings

according to a distance metric, such as cosine similarity. The core feature of this framework is a novel dual-backbone feature extractor, whose architecture is detailed in Figure 2. This model synergistically combines representations from two distinct foundation models: DINOv2 (15) and GastroNet (16), which were pretrained on a broad corpus of natural images and on large-scale endoscopic data (five million images), respectively. To create a highly discriminative embedding space, the entire architecture was optimized using a triplet loss-based metric learning approach.

2.1 Dual-backbone-based feature extraction network

The premise of the dual-backbone architecture is that a general-purpose model, such as DINOv2 (15), captures fundamental visual primitives (e.g., shapes and textures), while a domain-specific model, such as GastroNet (16), extracts the fine-grained features unique to endoscopic environments. DINOv2 enables strong geometric and semantic generalization across visual domains, whereas GastroNet provides domain-specific sensitivity to mucosal texture and color variations. Hence, we adopted DINOv2 as the general-purpose backbone in this study. By fusing these complementary representations within a Vision Transformer (ViT) architecture (17), the framework generates a more discriminative embedding space than either model could produce alone, enabling robust anatomical similarity matching for endoscopic image retrieval.

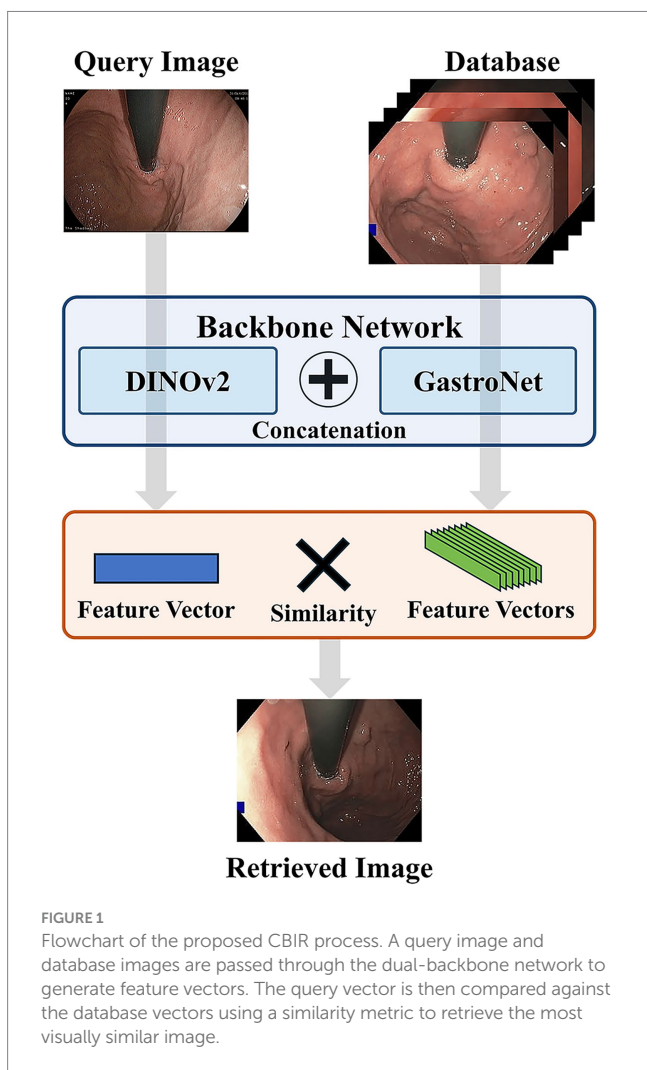
2.2 Feature pooling and fusion

Following feature extraction, the multi-dimensional feature maps from the dual backbones undergo a pooling and fusion process, which is detailed in the central column of Figure 2. This stage is engineered to distill the rich information from both backbones into a compact, discriminative representation. Generalized mean (GeM) pooling is employed to transform the 2D feature maps from each backbone into a 1D vector (18). This method enhances the overall representational power of the embedding while preserving important local features. The GeM pooling operation is defined in Equation (1) as:

$$f_k^{(g)} = \left(\frac{1}{|X_k|} \sum_{x \in X_k} x^{p_k} \right)^{\frac{1}{p_k}} \tag{1}$$

where the learnable parameter p_k controls the pooling behavior. Setting $p_k = 1$ is equivalent to average pooling, whereas $p_k \rightarrow \infty$ corresponds to max pooling.

To create a comprehensive feature vector that leverages the strengths of both models, the 1D vectors from each backbone were fused via concatenation, as illustrated in Figure 2. This concatenated embedding simultaneously captures general-purpose visual structures and domain-specific endoscopic details in a single, unified representation. While attention-based or learnable weighting mechanisms can also enable adaptive feature fusion, we adopted simple concatenation to ensure architectural simplicity and training stability. This approach preserves complementary information from



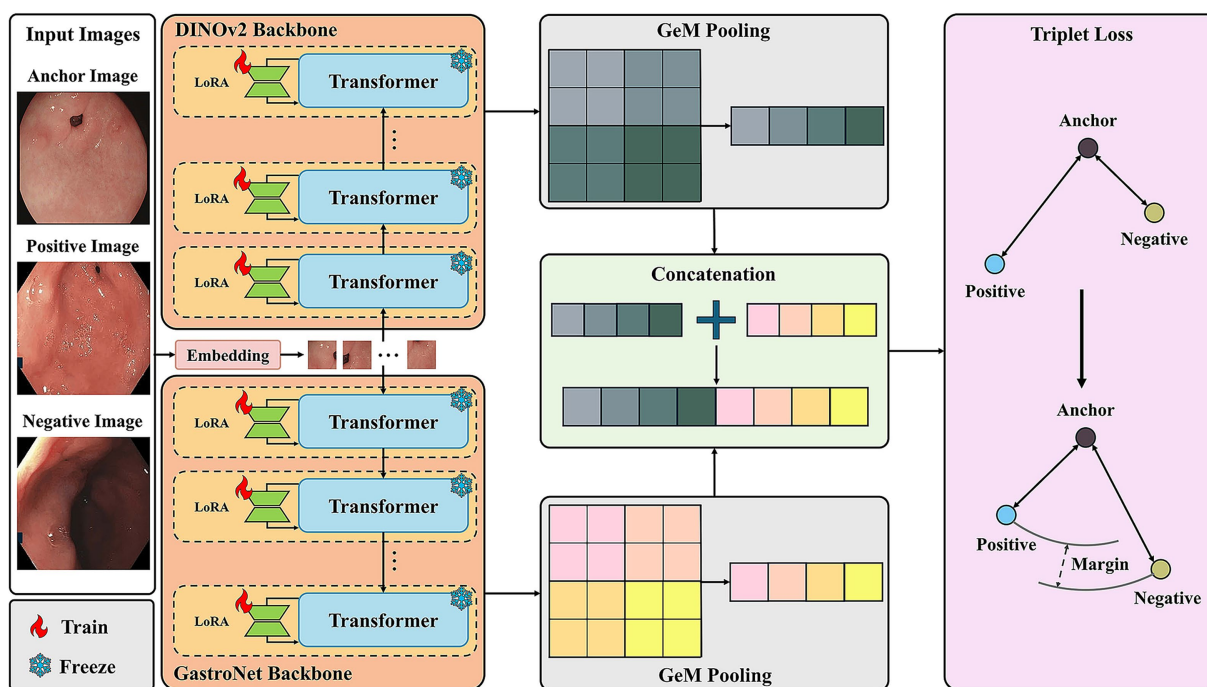


FIGURE 2 Architecture of the dual-backbone image-retrieval framework. Input images (anchor, positive, and negative) are processed by two parallel backbones: DINOv2 and GastroNet. PEFT is applied using LoRA modules, where pretrained transformer weights are frozen. Features from each backbone are pooled using generalized mean (GeM) pooling and then concatenated. The final embedding is optimized using a triplet loss function to minimize the distance between the anchor and positive while maximizing the distance to the negative.

both backbones without introducing additional parameters or alignment constraints, offering an efficient and widely applicable baseline for multimodal fusion (19, 20).

2.3 Model training

The similarity between two feature vectors is measured using cosine distance, which offers superior robustness to other metrics. The distance between two vectors, a and b , is calculated in Equation (2) as follows:

$$\text{dist}(a,b) = 1 - \frac{a \cdot b}{\|a\| \|b\|} \quad (2)$$

The embedding space was optimized for the retrieval task using a triplet loss function, which operates on data triplets comprising an anchor (x_a), a positive (x_p , an image similar to the anchor), and a negative (x_n , an image dissimilar to the anchor). The objective is to ensure that the anchor-positive distance is smaller than the anchor-negative distance by at least a predefined margin, α . This process guides the model to form distinct clusters of similar images within the embedding space, thereby optimizing its structure for retrieval. The loss is formally defined in Equation (3) as

$$L_{\text{triplet}} = \sum_{i=1}^N \max(d(f(x_a^i), f(x_p^i)) - d(f(x_a^i), f(x_n^i)) + \alpha, 0) \quad (3)$$

where $f(x)$ is the embedding generated by our dual-backbone model for an input image x , and $d(\cdot)$ is the cosine distance from Equation 2.

As illustrated in Figure 2, we adopted a parameter-efficient fine-tuning (PEFT) strategy using low-rank adaptation (LoRA) to minimize computational demands and the risk of overfitting (19–21).

2.4 Datasets

To validate model performance, we used four distinct datasets for training, in-domain evaluation, generalization testing, and synthetic querying. For model training and in-domain evaluation, we used two public datasets—Kvasir and HyperKvasir (22, 23)—which contain a substantial collection of endoscopic images categorized by anatomical landmarks and pathological findings. An endoscopic expert selected seven relevant classes, excluding cases involving artificial dyeing agents or ambiguous interpretations: esophagitis, normal pylorus, normal Z-line, ulcerative colitis, normal cecum, polyps, and retroflex stomach (Table 1). A balanced training dataset of 3,500 images was created by randomly sampling 500 images from each class. For the test set, 200 images were randomly sampled from the remaining images in each class.

To test the model with synthetic queries virtually generated at Soonchunhyang University, we used a dataset created by applying synthetic textures—derived from real endoscopic videos—to a 3D

TABLE 1 Distribution of images from the Kvasir and HyperKvasir datasets across the seven selected classes.

Dataset	Esophagitis	Normal pylorus	Normal Z-line	Ulcerative colitis	Normal cecum	Polyps	Retroflex stomach
Kvasir	1,000	1,000	1,000	1,000	1,000	1,000	—
HyperKvasir	663	999	932	851	1,009	1,028	764

model constructed from CT scans. These images were used exclusively as queries to assess performance on out-of-distribution data.

Finally, to assess generalizability to unseen clinical data, we employed the GastroHUN dataset (24), which contains an extensive collection of clinical endoscopic videos with anatomical labels. We created a comprehensive search database by sampling frames from these videos and used it to validate the robustness and clinical applicability of our model.

2.5 Evaluation metrics and parameter settings

Two standard retrieval metrics were adopted to evaluate model performance: Recall@k and mean average precision (mAP) (25–27). Recall@k measures the proportion of queries for which at least one correct image is retrieved within the top-k results. We adopted Recall@1 (R@1) to assess the models' ability to immediately find a relevant match and Recall@5 (R@5) to evaluate performance in a practical scenario where a clinician might review the top few suggestions.

mAP provides a more holistic evaluation of the ranked retrieval results. Unlike Recall@k, it considers the rank of all correct images in the retrieved list, rewarding models that place correct items higher and penalizing those that place correct items lower. When calculated over all queries, it provides a comprehensive, single-figure summary of a model's overall retrieval quality.

All models were implemented using the PyTorch framework and trained on a workstation with a single NVIDIA RTX 3090 GPU (24 GB VRAM). We employed the AdamW optimizer, a robust variant of Adam that improves regularization by decoupling weight decay from the gradient update (28). A learning rate of 1×10^{-5} was used, as this is a standard choice that facilitates stable convergence when fine-tuning large pretrained models. The model was trained for 30 epochs. The margin hyperparameter α in the triplet loss function was set to 0.3, a value selected to ensure sufficient separation between dissimilar classes without making the training excessively difficult.

3 Results

3.1 Image-retrieval results

To evaluate the effectiveness of our proposed model, we conducted comparative experiments against multiple representative baseline architectures. These baselines were organized into three categories: (1) commonly used convolutional neural network architectures pretrained on ImageNet (29) [ResNet50 (30), VGG19 (31), DenseNet

TABLE 2 Image-retrieval performance comparison of various models on the Kvasir and HyperKvasir test sets.

Model	Published in	Recall@1 (%)	Recall@5 (%)	mAP (%)
ResNet50	CVPR 2016	76.57	92.86	51.33
VGG19	ICLR 2015	75.71	92.57	49.17
DenseNet	CVPR 2017	80.86	90.23	51.77
SENet	CVPR 2018	81.14	91.38	51.48
ViT	ICLR 2021	82.86	93.35	45.07
Swin Transformer	ICCV 2021	81.71	91.46	53.75
DINOv1	ICCV 2021	83.71	93.29	60.36
DINOv2	TMLR 2023	86.86	93.43	71.73
GastroNet	MIA 2024	90.57	94.57	83.19
Ours		97.71	99.14	96.74

mAP, mean average precision; CVPR, Computer Vision and Pattern Recognition; ICLR, International Conference on Learning Representations; ICCV, International Conference on Computer Vision; TMLR, Transactions on Machine Learning Research; MIA, Medical Image Analysis. Bold values indicate the performance of the proposed dual-backbone model (Ours).

(32), and SENet (33)]; (2) supervised Transformer models [ViT-L/16 (17) and Swin Transformer (34)]; and (3) the foundation models that are direct components of our architecture [DINOv1 (35), DINOv2 (15), and GastroNet (16)]. For a fair and rigorous comparison, all baseline models were fine-tuned on our training set under identical experimental conditions. The quantitative results for the Kvasir and HyperKvasir test sets are presented in Table 2.

Our model achieved state-of-the-art performance across all metrics, attaining a Recall@1 of 97.71%, Recall@5 of 99.14%, and mAP of 96.74%. This performance significantly surpasses all baselines, including the performance of the strongest single-backbone model, GastroNet (Recall@1 = 90.57%, mAP = 83.19%). Notably, our dual-backbone fusion strategy delivers a 7.14 percentage point improvement in Recall@1 over GastroNet and 13.55% improvement in mAP. Figure 3 presents a qualitative comparison of retrieval results for four representative examples. The figure, which compares our model's top-three retrievals against those of key baseline models, visually confirms the superior quality of our model in consistently identifying semantically and visually coherent results. For esophagitis queries, our model retrieves images that match the inflammatory patterns and severity, whereas baselines such as the Swin Transformer incorrectly retrieve images of healthy esophageal tissue. Similarly, for viewpoint-sensitive queries such as normal pylorus, our model correctly identifies the precise endoscopic orientation (e.g., distal pyloric view). By contrast, even strong baselines such as DINOv1 retrieve the correct organ but fail to match the required perspective. These qualitative

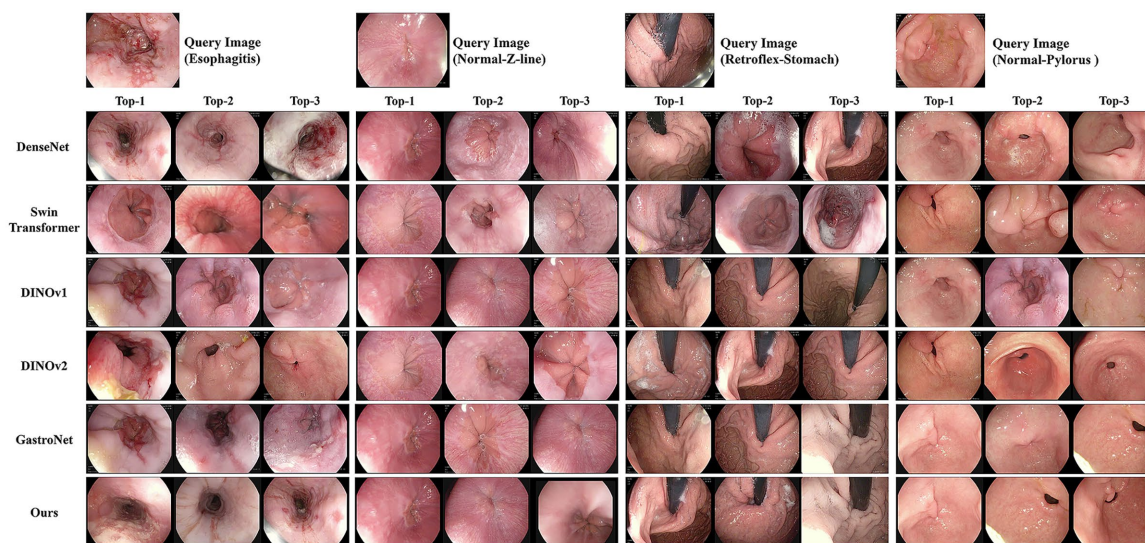


FIGURE 3 Qualitative comparison of image-retrieval results from the proposed model and representative baselines. The four columns show the retrieval results for four different query images: esophagitis, normal Z-line, retroflex-stomach, and normal-pylorus. The rows display the top-three most similar images retrieved by each model, allowing for a visual assessment of their performance in identifying relevant anatomical landmarks and pathological conditions.

TABLE 3 Class-wise image-retrieval performance on the Kvasir and HyperKvasir test set.

Class	DINOv2			GastroNet			Ours		
	Recall@1 (%)	Recall@5 (%)	mAP (%)	Recall@1 (%)	Recall@5 (%)	mAP (%)	Recall@1 (%)	Recall@5 (%)	mAP (%)
Esophagitis	84.00	90.00	69.85	88.00	90.00	79.84	96.00	98.00	89.76
Normal pylorus	92.00	98.00	76.32	92.00	100.00	86.16	100.00	100.00	99.99
Normal Z-line	86.00	96.00	68.48	90.00	96.00	83.47	98.00	100.00	98.46
Retroflex stomach	94.00	98.00	93.38	94.00	100.00	96.71	100.00	100.00	100.00
Ulcerative colitis	78.00	88.00	61.47	90.00	92.00	76.64	96.00	98.00	93.80
Normal cecum	88.00	90.00	71.67	88.00	90.00	78.92	96.00	98.00	96.88
Polyps	86.00	94.00	60.94	92.00	94.00	80.62	98.00	100.00	98.30

mAP, mean average precision. Bold values indicate the performance of the proposed dual-backbone model (Ours).

results corroborate the quantitative data presented in Table 2, demonstrating the robustness of our dual-backbone architecture.

A detailed class-wise evaluation is presented in Table 3, further highlighting the model’s robust performance. The model demonstrates exceptional accuracy in identifying anatomical landmarks, achieving a Recall@1 of over 98% and Recall@5 of 100% for normal-pylorus, normal-Z-line, retroflex-stomach classes. The mAP was also high, particularly for the normal pylorus (99.99%), normal Z-line (98.46%), and retroflex stomach (100%) classes, while remaining strong for pathological findings such as Esophagitis (89.76%).

Additionally, we evaluated zero-shot generalization by using the unseen GastroHUN dataset as the search database, with queries drawn from both the public Kvasir/HyperKvasir datasets and a proprietary synthetic dataset. This configuration rigorously assessed the model’s capacity to bridge the domain gap between diverse query sources and a real-world clinical database. As shown in Figure 4, this

evaluation yielded three key findings. First, the pretrained GastroNet significantly outperforms DINOv2, highlighting the benefit of endoscopic-specific pretraining. Second, triplet loss fine-tuning substantially improves the performance of all models. Finally and most critically, our dual-backbone architecture consistently outperforms all other configurations by retrieving semantically precise matches, demonstrating its superior robustness and generalizability.

To further examine the trade-off between retrieval accuracy and computational efficiency, we measured the average inference time per image on both GPU and CPU for the representative models (DINOv2, GastroNet, and ours). As summarized in Table 4, our dual-backbone model achieved the highest retrieval accuracy with moderate computational overhead (15.47 ms on GPU and 58.2 ms on CPU). Compared with the fastest single-backbone model (GastroNet, 10.91 ms on GPU and 26.5 ms on CPU), our model shows approximately 1.4 × higher GPU latency and 2.2 × higher CPU latency, reflecting the additional computation from dual-branch

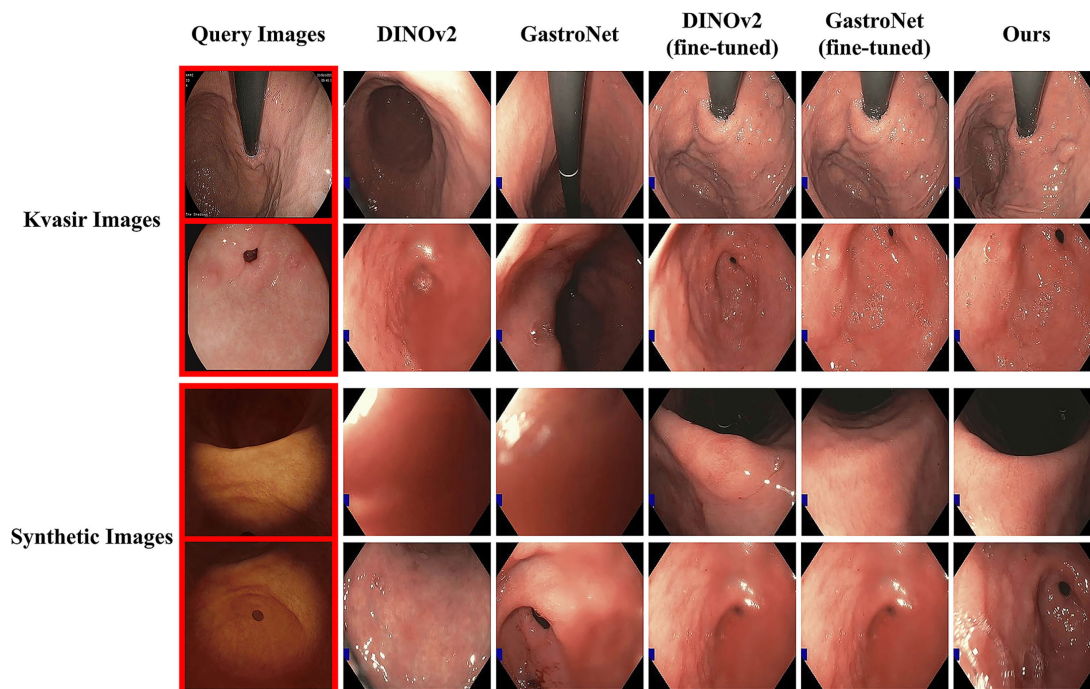


FIGURE 4 Qualitative comparison of zero-shot generalization performance. The images on the first column are selected as query images from the Kvasir dataset (top two rows) and a synthetic dataset (bottom two rows). The rightmost columns display the top-retrieved image from five different model configurations.

TABLE 4 GPU and CPU inference latency comparison of representative models (DINOv2, GastroNet, and our model).

Model	Recall@1 (%)	Recall@5 (%)	mAP (%)	GPU latency (ms)	CPU latency (ms)
DINOv2	86.86	93.43	71.73	10.60	34.2
GastroNet	90.57	94.57	83.19	10.91	26.5
Our model	97.71	99.14	96.74	15.47	58.2

mAP, mean average precision.

fusion. Nevertheless, the performance gain (Recall@1, 97.71% vs. 90.57%) demonstrates a favorable balance between accuracy and efficiency. Latency was measured on an NVIDIA RTX 3090 (24 GB VRAM) and an Intel Core i7-10700F CPU (2.90 GHz, 8 cores, 16 threads).

3.2 Ablation study

An ablation study was conducted to isolate the contribution of each component, and the results are summarized in Table 5. The dual-backbone fusion (GastroNet + DINOv2) achieved an mAP of 96.74%, outperforming both the single-backbone GastroNet (96.26%) and DINOv2 (76.57%), thereby confirming the synergistic effect of combining domain-specific and general features. Furthermore, GeM pooling (96.74% mAP) significantly outperformed average (84.05%) and max (89.75%) pooling. As shown in Figure 5, GeM pooling demonstrated superior visual coherence in its retrievals. Finally, triplet loss (96.74% mAP) outperformed contrastive loss (91.58%), validating its efficacy in structuring the embedding space for fine-grained retrieval.

3.3 Clinical effectiveness of the dual-backbone image retrieval model

To verify the effective operation of our dual-backbone image retrieval model across diverse clinical endoscopy videos, we conducted a query using actual esophagogastroduodenoscopy videos that accurately captured the four anatomical sites of stomach recommended by clinical guidelines: cardia, angle, body, and antrum (1). For each query, we extracted the top-one retrieved frame from the GastroHUN dataset videos and had the results reviewed by a clinical endoscopy specialist to confirm its accuracy. Our model successfully retrieved the correct images for all four targeted observation sites. The results from five representative videos are shown in Figure 6.

4 Discussion

Accurate photo documentation is a critical aspect of gastrointestinal endoscopy and serves as a significant quality indicator for examinations (2, 3). During EGD in particular, capturing images

TABLE 5 Ablation study on the effect of dual-backbone fusion.

DINOv2	GastroNet	Fusion	Average pooling	Max pooling	GeM pooling	Contrastive loss	Triplet loss	Recall@1 (%)	Recall@5 (%)	mAP (%)
✓					✓		✓	87.71	94.43	76.57
	✓				✓		✓	91.14	95.71	96.26
		✓	✓				✓	90.57	96.63	84.05
		✓		✓			✓	93.15	96.29	89.75
		✓			✓	✓		95.71	96.29	91.58
		✓			✓		✓	97.71	99.14	96.74

mAP, mean average precision. Bold values indicate the performance of the proposed dual-backbone model (Ours).

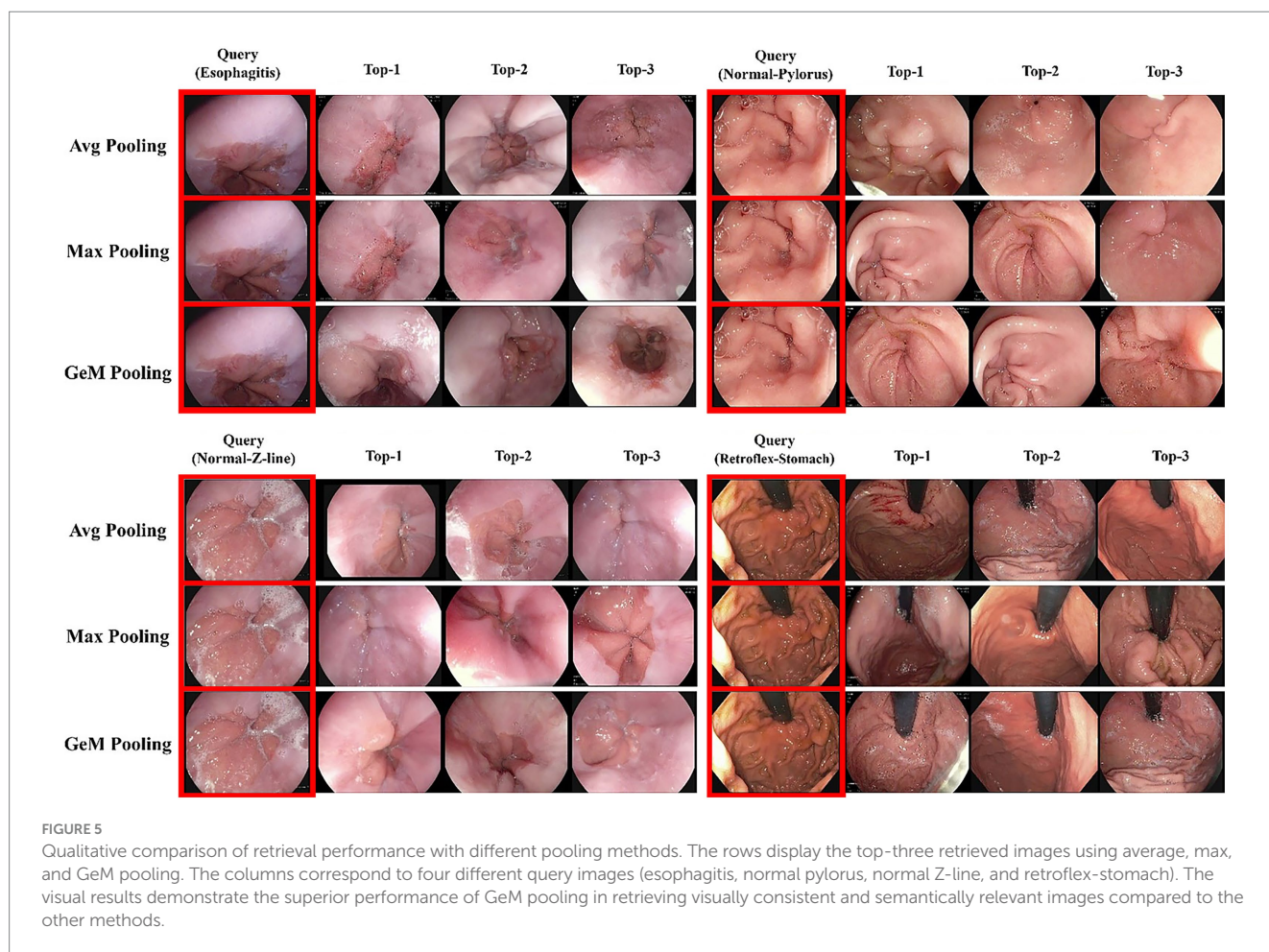


FIGURE 5 Qualitative comparison of retrieval performance with different pooling methods. The rows display the top-three retrieved images using average, max, and GeM pooling. The columns correspond to four different query images (esophagitis, normal pylorus, normal Z-line, and retroflex-stomach). The visual results demonstrate the superior performance of GeM pooling in retrieving visually consistent and semantically relevant images compared to the other methods.

of essential observation sites is strongly emphasized (1). In Korea, medical institutions evaluate the standard management of gastrointestinal endoscopy by verifying whether photographs of the recommended observation sites are correctly documented in randomly selected cases (36). However, manually selecting and verifying these images is extremely labor-intensive. The application of deep learning to automate this process has yielded impressive results. For instance, Choi et al. (4) introduced a multiclass classification system to recognize eight landmarks in the pictorial results of EGD, achieving an accuracy of 97.58%. Their study used 2,599 images captured from 250 participants using a specific Olympus CV-290 endoscope system. Similarly, Ahn et al. (5) developed an automated tool to capture 11 landmark images from endoscopic videos, achieving

98.16% accuracy; for their model, 102,798 photos from 3,309 examinations were used for training and validation. Despite these excellent results, their performance cannot be assured across various endoscopic models because the training data were derived from a limited range of hardware systems (8). Furthermore, the need for such a large, expertly annotated dataset complicates the development and scalability of these deep-learning models.

Our dual-backbone model was designed to overcome the limitations of these methods, and it demonstrated superior image-retrieval performance. This efficacy was confirmed not only quantitatively on a modest public dataset of 3,500 images but also through its robust performance on entirely unseen real-world and synthetic data. The architecture is built on the core hypothesis that

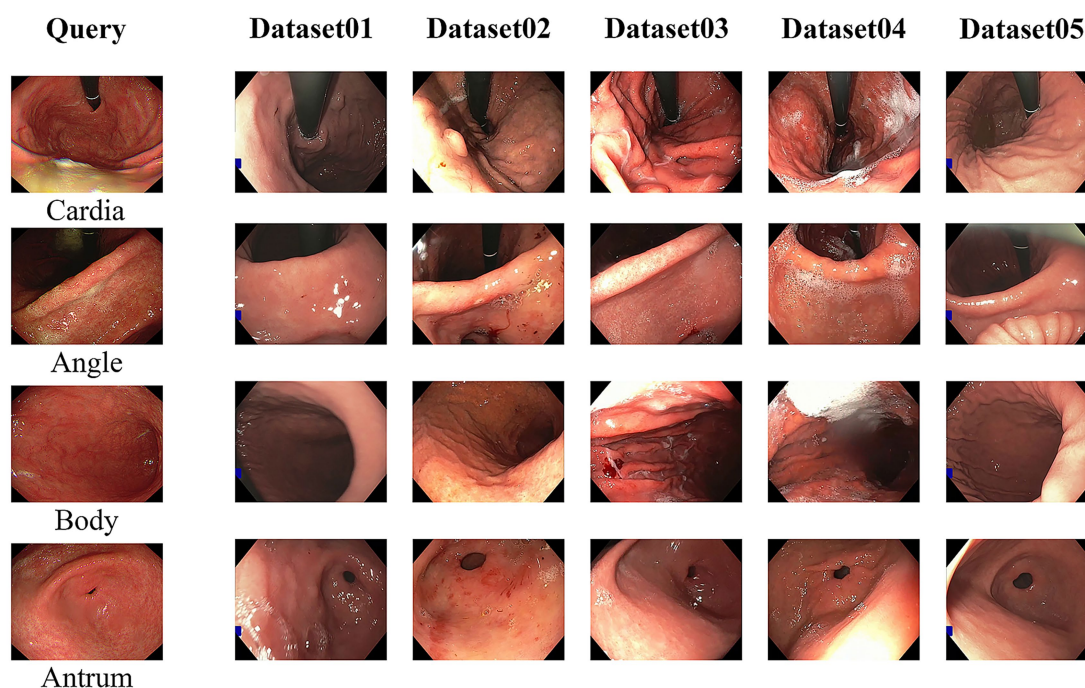


FIGURE 6

Image retrieval results for the recommended stomach observation sites. Using the four recommended stomach observation site images as queries, our dual-backbone-based model successfully extracted images from the GastroHUN video dataset at a clinically satisfactory level.

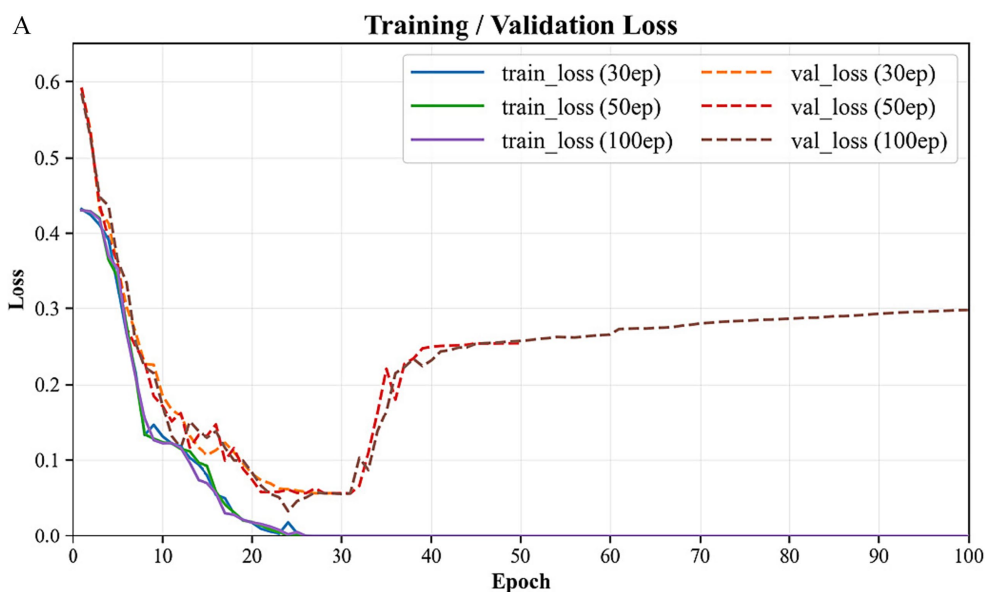
synergy can be achieved by combining two distinct models: a general-purpose model that captures fundamental visual primitives (e.g., shapes, textures, and spatial relationships) and a domain-specific model that discerns fine-grained features unique to the endoscopic environment. To achieve this, we used DINOv2, which was pretrained on a large-scale natural image dataset using a self-supervised learning scheme, to provide a foundation understanding of visual scenes and ensure robustness. This was integrated with its domain-specific counterpart, GastroNet, a foundation model pretrained on a large dataset of approximately 5 million endoscopic images. Both backbones leverage the ViT architecture (17), processing images by partitioning them into sequences of fixed-size patches and projecting them into high-dimensional embeddings. The fusion of these complementary representations ultimately yielded a more robust and discriminative embedding than either model could achieve independently. This dual-backbone approach aligns with recent research highlighting the importance of domain-specific feature learning for abnormality detection and the integration of self-supervised pretrained foundations with domain-specific models for mitigating data scarcity and improving robustness to clinical variability in medical imaging (37, 38).

To further analyze this synergistic interaction, we conducted an additional experiment to examine how different training strategies affect feature alignment between the two backbones. Specifically, we implemented a staged training strategy in which the GastroNet branch was frozen while only the DINOv2 branch was trained, followed by the joint fine-tuning of both backbones. We then tested two frozen-stage settings—7 epochs (mAP 95.86%, Recall@1 97.43%) and 10 epochs (mAP 96.58%, Recall@1 97.71%)—and found that both yielded slightly worse results than the dual-backbone baseline (mAP 96.74%, Recall@1 97.71%). These results indicate that simultaneous

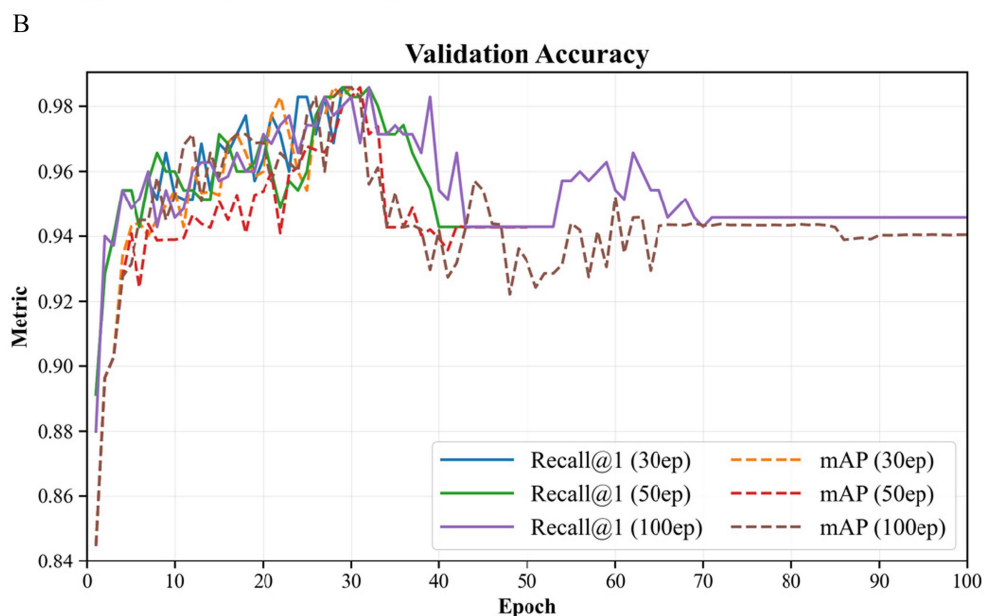
co-optimization enables more stable and efficient feature alignment between the two backbones compared with sequential adaptation. Therefore, we retained the joint fine-tuning strategy as the final configuration for all experiments.

To verify whether the model was sufficiently trained, we examined the training and validation loss, as well as validation accuracy (Recall@1 and mAP), across different epoch settings (30, 50, and 100 epochs). As shown in Figure 7, both loss and accuracy curves stabilized around the 25th to 30th epoch. The validation metrics peaked near 30 epochs, after which extended training caused a gradual rise in validation loss and a slight decrease in accuracy, indicating overfitting. Fluctuations in validation metrics were also observed, likely due to the inherent instability of triplet-based embedding learning under semi-hard negative sampling (39). Therefore, the optimal training duration for all final experiments was set as 30 epochs.

CBIR has emerged as a flexible and interpretable paradigm that overcomes the constraints of traditional deep-learning-based classification (10). Instead of assigning rigid labels, a CBIR system retrieves visually similar images from a reference database, an approach that allows for direct visual comparison, enhances clinical explainability, and accommodates images that defy predefined categories. Modern CBIR pipelines integrate three core components: feature extraction, similarity search, and metric learning. The process begins with feature extraction, where an image's visual content is converted into a quantitative vector representation, known as an embedding (40). Next, a similarity search compares the query embedding against the database; for large-scale applications, this is often accelerated using approximate nearest neighbor algorithms, such as Facebook AI Similarity Search, to avoid computationally prohibitive exhaustive searches (41). Finally, metric learning



ep, number of epochs for training the model



ep, number of epochs for training the model

FIGURE 7 Training and validation curves across different epoch settings (30, 50, and 100). Both (A) training and validation loss and (B) validation accuracy (Recall@1 and mAP) curves stabilize around the 25th–30th epoch, after which extensive training leads to overfitting, confirming that 30 epochs are sufficient for convergence.

techniques, such as triplet loss (13), optimize the quality of the embedding space by refining it to pull similar images closer together while pushing dissimilar ones apart. This integrated pipeline provides a scalable and interpretable framework uniquely suited to the complexity of endoscopic imaging. Our model’s training process and architecture were specifically tailored to the challenges of endoscopic imaging. Using a metric learning approach, triplets of anchor, positive, and negative images were input into the pretrained ViT-based backbones to generate embedding vectors (17). The ViT architecture, which partitions each image into patches before transformation, is

adept at capturing global context. We then employed GeM pooling (18), a critical component for this task. As a generalized form of both average and max pooling, GeM pooling excels at preserving local features while enhancing the embedding’s overall representational power. This capability proved particularly beneficial for endoscopic images, which are often characterized by subtle anatomical differences and complex textures, ultimately enabling the extraction of highly discriminative embeddings.

Our work aligns the recent paradigm shift in CBIR from traditional task-specific fine-tuning toward leveraging large-scale foundation models.

However, directly fine-tuning these massive models is computationally demanding and prone to overfitting. To mitigate these challenges, we employed a PEFT strategy using LoRA (21). This technique involves freezing the pretrained model weights and inserting small, trainable LoRA modules into each transformer block. By representing weight updates with low-rank matrices, LoRA drastically reduces the number of trainable parameters, i.e., from 43.68 million for full fine-tuning to just 0.615 million in our implementation. This approach yields significant advantages: it improves computational efficiency, reduces operational costs, and prevents catastrophic forgetting by preserving the model’s pretrained knowledge while skillfully adapting it to our retrieval task (42). Ultimately, PEFT strategies like LoRA enable the practical adaptation of powerful foundation models for specialized tasks with minimal trainable parameters.

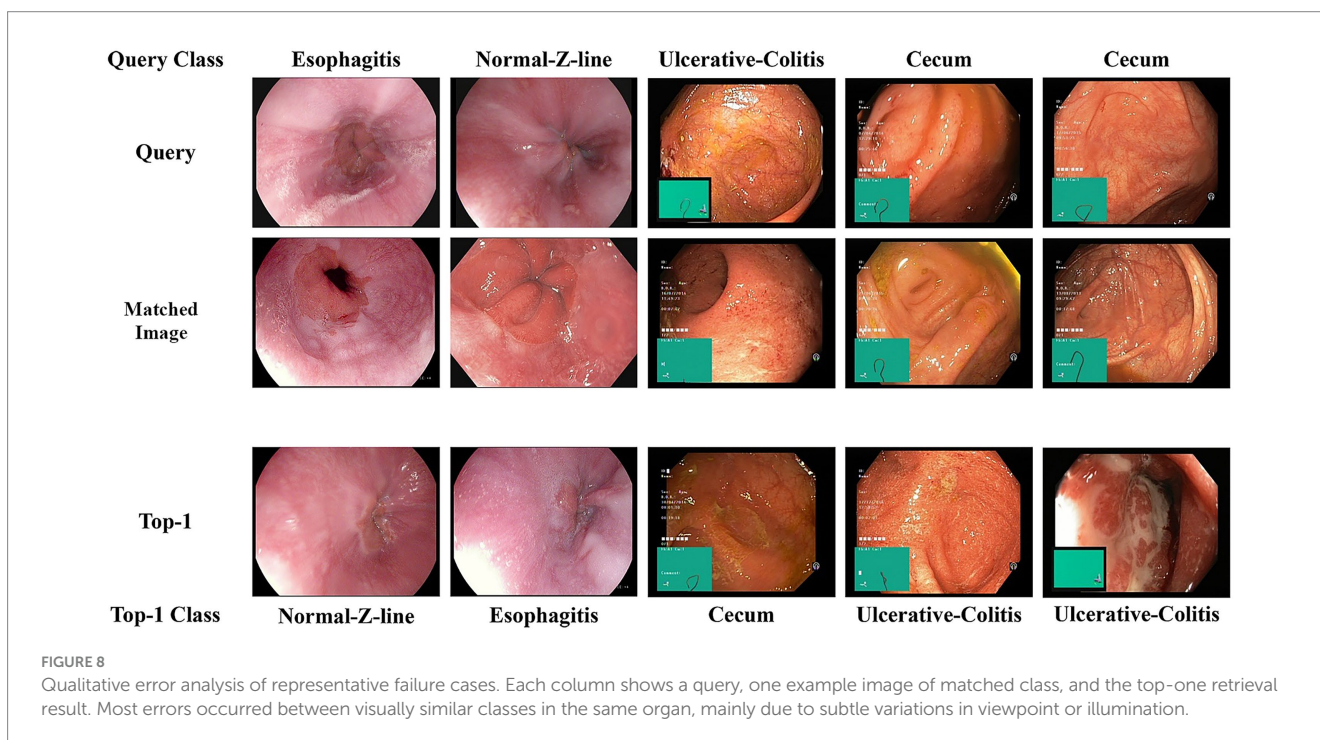
While foundation models offer robust representations that excel in few-shot and zero-shot scenarios (14), their direct application to medical data faces significant challenges. A major challenge is the pronounced domain shift between the natural images used for pretraining and the unique visual characteristics of medical data, which can degrade performance and necessitate sophisticated adaptation techniques (43, 44). Development is further constrained by strict privacy regulations and the high cost of expert annotation for data curation (45). Moreover, the inherent opacity of these “black box” algorithms can impede clinical adoption, where transparency is essential for building trust and ensuring patient safety (46). In light of these challenges, we acknowledge the limitations of our study. Although our model demonstrated strong performance, its robustness must be further validated on larger, multi-center datasets to confirm its generalizability across different clinical environments.

To contextualize these challenges, we analyzed representative retrieval failures. The qualitative results are shown in Figure 8. Most retrieval errors occurred between visually similar or anatomically adjacent categories (e.g., normal Z-line and esophagitis) and were primarily driven by subtle

variations in viewpoint or illumination rather than gross semantic confusion. These observations highlight the need for stronger fine-grained discrimination under variable imaging conditions. To further address these limitations, future work should extend the single-image retrieval framework to multi-image and multimodal settings, incorporating CLIP-based vision–language alignment to enhance semantic robustness (47, 48). Furthermore, more advanced and adaptive fusion strategies, such as attention-based or cross-modal mechanisms, should be investigated to further enhance feature interaction and retrieval robustness. Future work should also explore model compression and quantization to ensure its efficient deployment in real-time settings with constrained computational resources.

Although our dual-backbone architecture slightly increases inference latency (15.47 ms vs. 10.91 ms on GPU, $\approx 1.42\times$), this trade-off is considered acceptable given the substantial improvement in retrieval accuracy (+7.1% in Recall@1 and +13.6% in mAP). The proposed dual-backbone model therefore achieves a favorable balance between computational efficiency and accuracy. Additionally, incorporating an image quality assessment module could further enhance dataset consistency and improve the robustness and reliability of retrieval performance, which we plan to implement in future work.

In conclusion, we introduced a dual-backbone retrieval framework that establishes a new state-of-the-art for the automated quality control of endoscopic documentation. Our work demonstrates that the synergistic combination of a general-purpose and a domain-specific model yields a representation more powerful than either could achieve independently. By leveraging this architecture alongside triplet-loss-based metric learning, our approach surpasses traditional classification methods, offering superior explainability and the flexibility to manage ambiguous or novel visual data. Overall, this research contributes to the development of more effective and clinically applicable AI technologies for medical imaging.



The full implementation of the proposed framework is available at <https://github.com/Girin325/ImageRetrieval-with-DualModel>.

Data availability statement

The datasets presented in this article are not readily available because the images can be reconstructed, thereby compromising the privacy of the patients. Requests to access the datasets should be directed to JP, junspark@schmc.ac.kr.

Ethics statement

The studies involving humans were approved by Soonchunhyang University Hospital, Seoul. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

KK: Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. JP: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. SK: Data curation, Writing – original draft, Writing – review & editing. YH: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported

References

1. Rey, JF, and Lambert, RESGE Quality Assurance Committee. ESGE recommendations for quality control in gastrointestinal endoscopy: guidelines for image documentation in upper and lower GI endoscopy. *Endoscopy*. (2001) 33:901–3. doi: 10.1055/s-2001-42537
2. Bisschops, R, Areia, M, Coron, E, Dobru, D, Kaskas, B, Kuvaev, R, et al. Performance measures for upper gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative. *Endoscopy*. (2016) 48:843–64. doi: 10.1055/s-0042-113128
3. Yadlapati, R, Early, D, Iyer, PG, Morgan, DR, Sengupta, N, Sharma, P, et al. Quality indicators for upper GI endoscopy. *Am J Gastroenterol*. (2025) 120:290–312. doi: 10.14309/ajg.0000000000003252
4. Choi, SJ, Khan, MA, Choi, HS, Choo, J, Lee, JM, Kwon, S, et al. Development of artificial intelligence system for quality control of photo documentation in esophagogastroduodenoscopy. *Surg Endosc*. (2022) 36:57–65. doi: 10.1007/s00464-020-08236-6
5. Ahn, BY, Lee, J, Seol, J, Kim, JY, and Chung, H. Evaluation of an artificial intelligence-based system for real-time high-quality photodocumentation during esophagogastroduodenoscopy. *Sci Rep*. (2025) 15:4693. doi: 10.1038/s41598-024-83721-9
6. Kim, MJ, Kim, SH, Kim, SM, Nam, JH, Hwang, YB, and Lim, YJ. The advent of domain adaptation into artificial intelligence for gastrointestinal endoscopy and medical imaging. *Diagnostics*. (2023) 13:3023. doi: 10.3390/diagnostics13193023
7. Jin, Z, Gan, T, Wang, P, Fu, Z, Zhang, C, Yan, Q, et al. Deep learning for gastroscopic images: computer-aided techniques for clinicians. *Biomed Eng Online*. (2022) 21:12. doi: 10.1186/s12938-022-00979-8
8. Park, J, Hwang, Y, Kim, HG, Lee, JS, Kim, JO, Lee, TH, et al. Reduced detection rate of artificial intelligence in images obtained from untrained endoscope models and improvement using domain adaptation algorithm. *Front Med*. (2022) 9:1036974. doi: 10.3389/fmed.2022.1036974
9. Park, D, and Hwang, Y. Efficient image retrieval using hierarchical K-means clustering. *Sensors*. (2024) 24:2401. doi: 10.3390/s24082401
10. Trojancanec, K, Dimitrovski, I, and Loskovska, S. (2009). Content based image retrieval in medical applications: an improvement of the two-level architecture. *IEEE EUROCON 2009*. 118–121
11. Ruano, J, Gómez, M, Romero, E, and Manzanera, A. Leveraging a realistic synthetic database to learn shape-from-shading for estimating the colon depth in colonoscopy images. *Comput Med Imaging Graph*. (2024) 115:102390. doi: 10.1016/j.compmedimag.2024.102390

by the National Research Foundation of Korea (NRF) Grant (No. RS-2023-00211951) and Innovative Human Resource Development for Local Intellectualization Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) Grant (IITP-2025-RS-2020-II201462, 50%) funded by the Korean Government (MSIT).

Acknowledgments

The authors would like to thank Editage (www.editage.co.kr) for English language editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

12. Ye, M, Johns, E, Walter, B, Meining, A, and Yang, GZ. An image retrieval framework for real-time endoscopic image retargeting. *Int J Comput Assist Radiol Surg.* (2017) 12:1281–92. doi: 10.1007/s11548-017-1620-7
13. Schroff, F, Kalenichenko, D, and Philbin, J. (2015). FaceNet: a unified embedding for face recognition and clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 815–823
14. Wiggins, WF, and Tejani, AS. On the opportunities and risks of foundation models for natural language processing in radiology. *Radiol Artif Intell.* (2022) 4:e220119. doi: 10.1148/ryai.220119
15. Oquab, M, Darcet, T, Moutakanni, T, Vo, H, Szafraniec, M, Khalidov, V, et al. DINOv2: learning robust visual features without supervision. *arXiv [Preprint]*. (2023) 1–31. doi: 10.48550/arXiv.2304.07193
16. Boers, TGW, Fockens, KN, van der Putten, JA, Jaspers, TJM, Kusters, CHJ, Jukema, JB, et al. Foundation models in gastrointestinal endoscopic AI: impact of architecture, pre-training approach and data efficiency. *Med Image Anal.* (2024) 98:103298. doi: 10.1016/j.media.2024.103298
17. Dosovitskiy, A, Beyer, L, Kolesnikov, A, Weissenborn, D, Zhai, X, Unterthiner, T, et al. (2021). An image is worth 16×16 words: transformers for image recognition at scale. *arXiv*. Available online at: <https://doi.org/10.48550/arXiv.2010.11929>. [Epub ahead of preprint]
18. Radenović, F, Tolias, G, and Chum, O. Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans Pattern Anal Mach Intell.* (2019) 41:1655–68. doi: 10.1109/TPAMI.2018.2846566
19. Gao, J, Li, P, Chen, Z, and Zhang, J. A survey on deep learning for multimodal data fusion. *Neural Comput.* (2020) 32:829–64. doi: 10.1162/neco_a_01273
20. Khan, S, Naseer, M, Hayat, M, Zamir, SW, Shahbaz Khan, F, and Shah, M. Transformers in vision: a survey. *ACM Comput Surv.* (2022) 54:1–41. doi: 10.1145/3505244
21. Hu, EJ, Shen, Y, Wallis, P, Allen-Zhu, Z, Li, Y, Wang, S, et al. (2022). LoRA: low-rank adaptation of large language models. *arXiv*. Available online at: <https://doi.org/10.48550/arXiv.2106.09685>. [Epub ahead of preprint]
22. Borgli, H, Thambawita, V, Smedsrud, PH, Hicks, S, Jha, D, Eskeland, SL, et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci Data.* (2020) 7:283. doi: 10.1038/s41597-020-00622-y
23. Pogorelov, K, Randel, KR, Griwodz, C, Eskeland, SL, de Lange, T, Johansen, D, et al. (2017). KVASIR: a multi-class image dataset for computer aided gastrointestinal disease detection. *MMSys'17: Proceedings of the 8th ACM on Multimedia Systems Conference*. 164–169
24. Bravo, D, Frias, J, Vera, F, Trejos, J, Martínez, C, Gómez, M, et al. Gastrohun an endoscopy dataset of complete systematic screening protocol for the stomach. *Sci Data.* (2025) 12:102. doi: 10.1038/s41597-025-04401-5
25. Harman, DK. (1993). *The First Text Retrieval Conference (TREC-1)*. 500–207
26. Harman, D. Evaluation issues in information retrieval. *Inf Process Manag.* (1992) 28:439–40. doi: 10.1016/0306-4573(92)90001-G
27. Buckley, C, and Voorhees, EM. Evaluating evaluation measure stability. *ACM SIGIR Forum.* (2017) 51:235–42. doi: 10.1145/3130348.3130373
28. Loshchilov, I, and Hutter, F. (2019). Decoupled weight decay regularization. *arXiv*. Available online at: <https://doi.org/10.48550/arXiv.1711.05101>. [Epub ahead of preprint]
29. Deng, J, Dong, W, Socher, R, Li, L-J, Li, K, and Fei-Fei, L. (2009). ImageNet: a large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255
30. He, K, Zhang, X, Ren, S, and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778
31. Simonyan, K, and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv*. Available online at: <https://doi.org/10.48550/arXiv.1409.1556>. [Epub ahead of preprint]
32. Huang, G, Liu, Z, Van Der Maaten, L, and Weinberger, KQ. (2017). Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269
33. Hu, J, Shen, L, Albanie, S, Sun, G, and Wu, E. Squeeze-and-excitation networks. *IEEE Trans Pattern Anal Mach Intell.* (2020) 42:2011–23. doi: 10.1109/TPAMI.2019.2913372
34. Carion, N, Massa, F, Synnaeve, G, Usunier, N, Kirillov, A, and Zagoruyko, S. (2020). Hierarchical vision transformer using shifted windows. *European Conference on Computer Vision (ECCV)*. 213–229
35. Caron, M, Touvron, H, Misra, I, Jegou, H, Mairal, J, Bojanowski, P, et al. (2021). Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9630–9640
36. Min, JK, Cha, JM, Cho, YK, Kim, JH, Yoon, SM, Im, JP, et al. Revision of quality indicators for the endoscopy quality improvement program of the national cancer screening program in Korea. *Clin Endosc.* (2018) 51:239–52. doi: 10.5946/ce.2018.075
37. Iqbal, I, Walayat, K, Kakar, MU, and Ma, J. Automated identification of human gastrointestinal tract abnormalities based on deep convolutional neural network with endoscopic images. *Intell Syst Appl.* (2022) 16:200149. doi: 10.1016/j.iswa.2022.200149
38. Abbas, T, Linjawi, M, Iqbal, I, Alghushairy, O, Alsini, R, and Daud, A. Significance of unifying semi and self-supervision for the radical improvement of medical imaging: a collaborative research effort. *Biomed Signal Process Control.* (2026) 111:108391. doi: 10.1016/j.bspc.2025.108391
39. Wu, CY, Manmatha, R, Smola, AJ, and Krähenbühl, P. (2017). Sampling matters in deep embedding learning. *2017 IEEE International Conference on Computer Vision (ICCV)*. 2859–2867
40. Hu, H, Zheng, W, Zhang, X, Zhang, X, Liu, J, Hu, W, et al. Content-based gastric image retrieval using convolutional neural networks. *Int J Imaging Syst Technol.* (2021) 31:439–49. doi: 10.1002/ima.22470
41. Johnson, J, Douze, M, and Jegou, H. Billion-scale similarity search with GPUs. *IEEE Trans Big Data.* (2021) 7:535–47. doi: 10.1109/TBDATA.2019.2921572
42. Lialin, V, Deshpande, V, Yao, X, and Rumshisky, A. (2024). Scaling down to scale up: a guide to parameter-efficient fine-tuning. *arXiv*. Available online at: <https://doi.org/10.48550/arXiv.2303.15647>. [Epub ahead of preprint]
43. Li, Y, Ghahremani, M, and Wachinger, C. (2025). MedBridge: bridging foundation vision-language models to medical image diagnosis. *arXiv*. Available online at: <https://doi.org/10.48550/arXiv.2505.21698>. [Epub ahead of preprint]
44. Li, Y, Wu, Y, Lai, Y, Hu, M, and Yang, X. (2025). MedDINOv3: how to adapt vision foundation models for medical image segmentation?. *arXiv*. Available online at: <https://doi.org/10.48550/arXiv.2509.02379>. [Epub ahead of preprint]
45. van Veldhuizen, V, Botha, V, Lu, C, Cesur, ME, Lipman, KG, de Jong, ED, et al. (2025) Foundation models in medical imaging—a review and outlook. *arXiv*. Available online at: <https://doi.org/10.48550/arXiv.2506.09095>. [Epub ahead of preprint]
46. Wadden, JJ. Defining the undefinable: the black box problem in healthcare artificial intelligence. *J Med Ethics.* (2022) 48:764–8. doi: 10.1136/medethics-2021-107529
47. Akbacak, E, Aydin, M, and Aydin, MA. MLMQ-IR: multi-label multi-query image retrieval based on the variance of hamming distance. *Knowl Based Syst.* (2024) 283:111193. doi: 10.1016/j.knsys.2023.111193
48. Radford, A, Kim, JW, Hallacy, C, Ramesh, A, Goh, G, and Agarwal, SG. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 8748–8763