Oncology Advances **2026** vol. 4(2) | e00006
DOI: 10.14218/OnA.2026.00006

Original Article

Enhancing Cross-dataset Zero-shot Generalization in Colorectal Polyp Detection Using Vision-language Models

Zhanglu Hu[1#], Xiaodan Chen[1#], Mingjia Ma[1], Bohan Liang[1], Weidong Zhang[1,2*], Jing Zhang[3*] and Sichao Tian[1*]

[1]*Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China;* [2]*Department of Phytochemistry, School of Pharmacy, Second Military Medical University (Naval Medical University), Shanghai, China;* [3]*Macao University of Science and Technology, Macao, China*

Received: March 02, 2026 | **Revised:** April 09, 2026 | **Accepted:** May 06, 2026 | **Published online:** June 02, 2026

Abstract

Background and objectives: Colorectal polyp detection from endoscopic images is critical for the early diagnosis of colorectal cancer. However, traditional deep learning methods often suffer from limited generalization when deployed across datasets containing different polyp morphologies. This work aimed to investigate whether vision-language foundation models can facilitate zero-shot generalization across multiple polyp datasets without target-domain fine-tuning.

Methods: We introduced a zero-shot colorectal polyp detection framework based on Contrastive Language-Image Pretraining (CLIP) to improve cross-dataset detection performance. Key innovations include: (1) a background patch contrastive loss using pseudo-normal tissue patches to teach the model to distinguish normal mucosa from polyps; (2) attribute-enhanced text prompts that incorporate domain-specific descriptors of polyp appearance, improving the model's semantic generalization to novel polyp morphologies; and (3) an enhanced CLIP visual adapter with per-layer adaptive feature fusion and generalized mean pooling to capture multi-scale features for better polyp localization. During training, we use one annotated colorectal polyp dataset (e.g., CVC-ColonDB) to learn patch-level image-text correspondence. The model is then evaluated in a zero-shot manner on different polyp datasets (CVC-ClinicDB, Kvasir-SEG, and CVC-300), where we measure both pixel-level and image-level anomaly detection performance.

Results: The framework demonstrated robust zero-shot generalization on unseen test cohorts. Without any dataset-specific fine-tuning, the model achieved a mean pixel-level AUROC of 0.94 and a mean average precision of 0.81 across the 12 leave-one-dataset-out zero-shot transfer settings. In the CVC-ColonDB-source benchmark, the model achieved a mean Dice coefficient of 0.84 across CVC-ClinicDB, Kvasir-SEG, and CVC-300. This high level of performance was consistent across datasets with distinct visual characteristics, underscoring the ability of the model to detect diverse polyp morphologies that it had not been explicitly trained to recognize.

Conclusions: Our findings demonstrate that an anomaly-aware vision-language model significantly improves cross-dataset polyp detection generalization without requiring normal images for training. This multimodal strategy may facilitate the robust deployment of artificial intelligence-based colorectal screening systems by enabling reliable detection of diverse polyp morphologies across different clinical settings. Extension to non-polyp colorectal pathologies (e.g., ulcerative colitis and colorectal tumors) remains an important direction for future work, pending the availability of pixel-level annotated datasets for these lesion categories.

Introduction

Keywords: Colorectal polyps; Endoscopy; Gastrointestinal tract; Deep learning; Natural language processing; Image interpretation; Computer-assisted diagnosis; Reproducibility of results.

***Correspondence to:** Weidong Zhang, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100193, China; Department of Phytochemistry, School of Pharmacy, Second Military Medical University (Naval Medical University), Shanghai 200433, China. ORCID: https://orcid.org/0000-0002-7384-2490. Tel: +86-57833013, E-mail: wdzhang@implad.ac.cn; Jing Zhang, Macao University of Science and Technology, Macao 999078, China. ORCID: https://orcid.org/0000-0001-9101-451X. Tel: +853-62358295, E-mail: jingzhang@must.edu.mo; Sichao Tian, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100193, China. ORCID: https://orcid.org/0000-0003-4887-368X. Tel: +86-17620951080, E-mail: sctian@implad.ac.cn

How to cite this article: Hu Z, Chen X, Ma M, Liang B, Zhang W, Zhang J, *et al*. Enhancing Cross-dataset Zero-shot Generalization in Colorectal Polyp Detection Using Vision-language Models. *Oncol Adv* 2026;4(2):e00006. DOI: https://doi.org/10.14218/OnA.2026.00006.

Colonoscopy is an effective tool for colorectal cancer prevention, enabling the detection and removal of precancerous polyps. Early detection and removal of colorectal polyps can significantly reduce colorectal cancer incidence. However, colonoscopy is a complex procedure, and even experienced endoscopists may miss a substantial fraction of polyps; a recent meta-analysis reported miss rates of approximately 26% for adenomatous polyps.[1] To assist clinicians, computer-aided detection (CADe) systems using artificial intelligence (AI) have been developed to help identify polyps.[2] While AI-driven systems have shown promising results in adenoma detection, their clinical utility is often limited by a significant challenge: poor generalization across different types of lesions. Most current CADe models are trained on large annotated datasets and tend to perform well only on lesion types similar to their training data. Their performance degrades substantially when encountering unseen lesion types, such as those with different

shapes, sizes, or textures.[3,4] Parallel advances likewise show that colorectal AI is expanding from real-time colonoscopic recognition and digital pathology quantification to machine-learning-based blood or laboratory screening and AI-supported malignant polyp characterization and surveillance,[5-10] collectively underscoring the need for robust systems that can transfer across heterogeneous clinical settings.

In this study, we investigate the issue of cross-dataset generalization, specifically a zero-shot lesion detection scenario across different types of colonic lesions. We aimed to train a generalizable model on one category of colorectal lesion and then enable it to detect other lesion types without additional training. For example, a model trained only on images of adenomatous polyps should also be able to identify a different lesion type, such as an early cancerous lesion, during testing. This "zero-shot" setting requires the model to recognize patterns from a class it has never explicitly seen during training, a capability that is essential for real-world clinical deployment, where lesion presentations are highly variable.[11,12]

Recent advances in vision-language models offer a potential solution.[13-15] From a translational gastroenterology perspective, recent reviews have further emphasized that advanced endoscopic imaging, optical biopsy, and lesion-tailored resection strategies are central to improving colorectal cancer detection and management.[16,17] A notable study introduced Anomaly-Aware CLIP (AA-CLIP), built upon Contrastive Language-Image Pretraining (CLIP) adapting the powerful CLIP model for industrial anomaly detection.[13,18] AA-CLIP enhances CLIP's ability to distinguish between normal and abnormal samples by learning anomaly-focused text "anchors" and aligning patch-level visual features to these anchors using transformer-based adapters. Building on the AA-CLIP architecture, we present a zero-shot lesion detection framework designed to overcome the cross-lesion detection problem. Our approach introduces a set of modifications to enhance its applicability to medical imaging.[19-24]

To improve the model's discrimination between pathologic and normal tissue, we develop a contrastive learning strategy using pseudo-normal image patches extracted from lesion-containing images. To enhance semantic understanding, we design attribute-enriched text prompts that describe the visual appearance of lesions using generalizable textual attributes. Finally, to further improve lesion localization, we refine the visual encoder by integrating residual adapters at multiple layers and applying generalized mean (GeM) pooling for superior multi-scale feature aggregation.[19-21,24-27]

We validate our model under a cross-lesion testing protocol: the model is trained on labeled images from a single dataset and subsequently evaluated on multiple unseen colonoscopy datasets containing a wide range of polyp pathologies, without using any target-domain data for model tuning. This experimental design simulates real-world deployment, where a model encounters data from unseen clinical scenarios, and directly tests the framework's generalization performance. This study aimed to develop and validate a generalizable AI framework for polyp detection across unseen colonoscopy datasets and to evaluate its performance under a cross-lesion testing protocol that simulates real-world clinical scenarios.

Materials and methods

Dataset and preprocessing

We evaluated the zero-shot lesion detection framework using four publicly available colonoscopy image datasets: CVC-ClinicDB, CVC-ColonDB, Kvasir-SEG, and CVC-300. CVC-ClinicDB contains 612 colonoscopy frames extracted from clinical videos, each annotated with expert-provided segmentation masks. CVC-ColonDB comprises publicly available colonoscopy images with corresponding pixel-level polyp annotations. Kvasir-SEG provides a larger public dataset of polyp images with pixel-level masks. The CVC-300 test subset used in this study contains 60 colonoscopy images with ground-truth polyp segmentation masks.

All images were resized to 518 × 518 pixels to match the input resolution of the Vision Transformer. We normalized the images using the standard mean and standard deviation from the original CLIP preprocessing. We adopted a rigorous leave-one-dataset-out evaluation protocol. For each cross-dataset experiment, images from one dataset were used for training (the source domain), and the trained model was then tested on the remaining three datasets (the target domains) in a zero-shot setting (with no fine-tuning on the target data). This protocol evaluates the ability of the model to generalize to unseen polyp pathologies and imaging conditions.

Model architecture

Our approach extends the AA-CLIP framework, which adapts the CLIP model for anomaly detection. CLIP consists of an image encoder, specifically a ViT-L/14 Vision Transformer, and a transformer-based text encoder to map images and text descriptions into a shared embedding space.[13] Although CLIP has shown strong zero-shot performance in general vision tasks, its standard formulation is anomaly-unaware, as it cannot inherently distinguish normal from abnormal features in fine-grained medical images. AA-CLIP is designed to address this gap by introducing lightweight transformer adapter modules and an anomaly-focused training strategy.[18-22,24] These adapters refine the model features for anomaly detection without disrupting the foundational knowledge acquired during CLIP pretraining.

The training process is structured in two sequential stages designed to align visual features with anomaly-specific textual concepts, as illustrated in Figure 1. In Stage 1, the focus is on refining the textual representations. The pre-trained CLIP text encoder is fine-tuned with lightweight adapters to produce distinct textual anchors for normal and abnormal semantics. This process produces a pair of text embeddings representing contrasting concepts: one for a normal description (e.g., healthy colon tissue) and one for an abnormal description (e.g., a colorectal lesion).

In Stage 2, the visual encoder is adapted to align with these optimized text anchors. Visual adapters are utilized to fine-tune the image encoder such that image patch features from anomalous regions align more closely with the abnormal text anchor than with the normal one. The model learns the correspondence between local, patch-level visual features and the text embeddings in the shared latent space. For inference, the model generates a pixel-wise similarity map by comparing extracted visual features with both normal and abnormal text anchors, thereby enabling localization of anomalous regions by identifying patches that exhibit higher similarity to the abnormal descriptions.

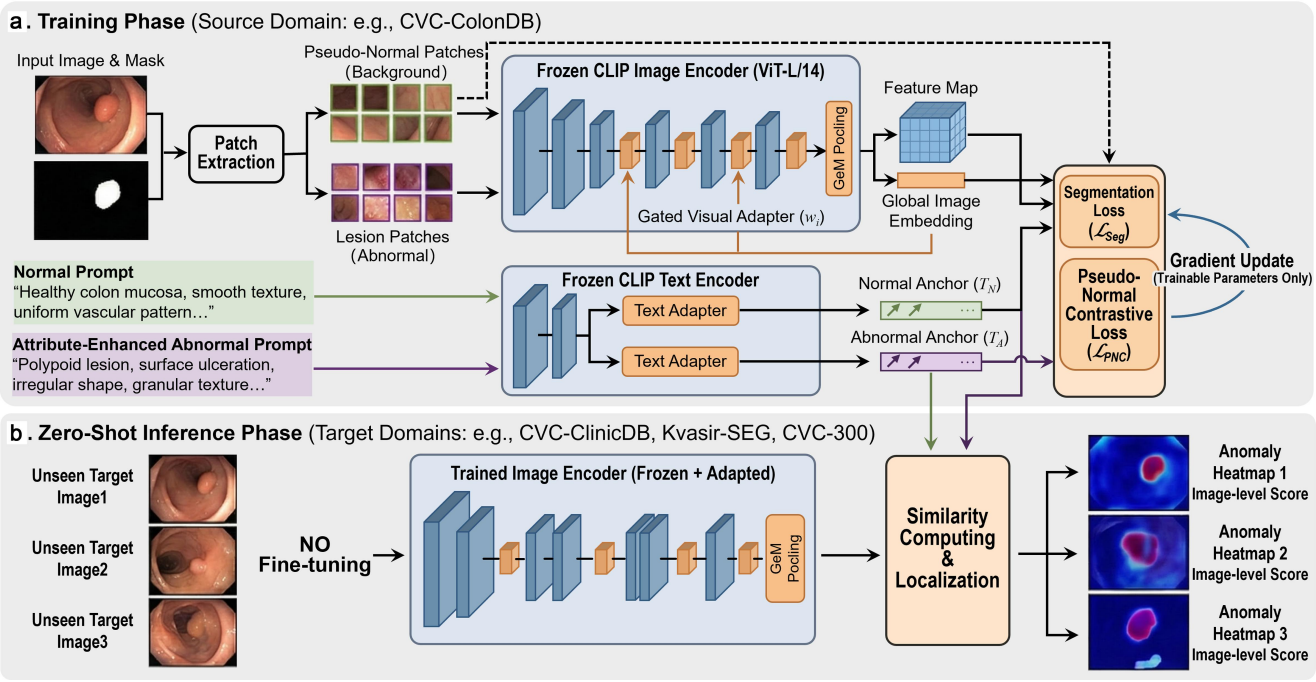

Fig. 1. Overview of the proposed zero-shot colorectal polyp detection framework. The framework leverages a frozen Contrastive Language-Image Pretraining (CLIP) backbone adapted via parameter-efficient modules. (a) Training Phase (Source Domain). Utilizing a single annotated dataset, Attribute-Enhanced Prompts guide the Text Encoder to generate discriminative normal and abnormal anchors. The Image Encoder is refined using Gated Visual Adapters and Generalized Mean (GeM) Pooling to capture multi-scale features. Training is guided by a segmentation loss (L_{Seg}) and a pseudo-normal contrastive loss (L_{PNC}), which uses background regions as pseudo-normal samples to distinguish lesions. Only the adapters are updated to preserve pre-trained knowledge. (b) Zero-Shot Inference Phase (Target Domains). The trained model is evaluated directly on unseen datasets without fine-tuning. Anomaly localization is achieved by computing pixel-wise similarity between the adapted visual features and the learned text anchors, generating precise anomaly heatmaps.

Medical attribute-enriched text prompts

The development of discriminative text prompts is crucial for effective zero-shot detection in vision-language models. Whereas typical approaches use simple prompts (e.g., "a photo of a diseased colon" versus "a photo of a normal colon"), we introduce medical attribute-enriched prompts that incorporate domain-specific descriptors of colon pathologies. We design a comprehensive set of text prompts detailing salient visual characteristics of lesions, moving beyond generic disease labels. For example, an abnormal prompt might be "colon mucosa with ulceration and irregular polypoid lesion" rather than "diseased colon." These medical-specific descriptions explicitly include visual attributes such as ulceration, bleeding, erosion, or polypoid growth.

This attribute-enriched approach provides a more nuanced semantic grounding for abnormality, leading to more descriptive and discriminative text anchors. Furthermore, because these attributes are generalizable across diverse lesion morphologies, this strategy improves cross-dataset generalization by preventing the model from overfitting to the dataset-specific visual features of a single lesion type. The abnormal text anchor becomes an attribute-aware representation of a lesion, which enhances its effectiveness when detecting novel lesion appearances in unseen test datasets.[20,21,24] The full set of prompts is provided in Supplementary Table 1.

Visual adapter refinement and feature aggregation

We propose a lightweight visual adapter strategy that modulates the CLIP ViT-L/14 image encoder with per-layer gating and a GeM pooling layer, designed to enhance cross-dataset generalization without introducing computationally expensive attention blocks.

First, we introduce adaptive per-layer gating for the visual adapters. While the original AA-CLIP design inserts residual adapters primarily in the early layers of the image encoder, we extend this approach by learning a scalar gating parameter ω_i^l for each adapter at layer l. This parameter scales the adapter's output before it is integrated into the main network via a residual connection:

$$\boldsymbol{F}_{out}^{(l)} = \boldsymbol{F}_{in}^{(l)} + \omega_i^{(l)} \cdot Adapter\left(\boldsymbol{F}_{in}^{(l)}\right)$$

These gating weights are learned during training, allowing the model to adaptively modulate the contribution of each adapter layer, optimizing the balance between pretrained knowledge and task-specific feature refinement.

Second, to derive a global image representation that complements the patch-level anomaly map, we apply GeM pooling over the vision encoder's output feature map. GeM

pooling is a differentiable operation that generalizes average and max pooling by taking the p-th power mean of feature activations, with p being a learnable parameter. This enables the model to dynamically prioritize regions with high activation while still aggregating information from the entire image, resulting in a robust global embedding summarizing the image's abnormal content. We use this pooled embedding for image-level anomaly classification in conjunction with the patch-level segmentation output. By refining patch features at multiple network layers and then pooling them, the model better captures both the fine-grained details and the global context necessary for accurate lesion localization.[19,28]

Contrastive learning with pseudo-normal patches

In the original AA-CLIP framework, image-level anomaly detection is guided by a binary cross-entropy classification objective. We hypothesize that a patch-level training loss better exploits local lesion cues and improves localization granularity. To this end, we replace the image-level loss with a pseudo-normal patch contrastive (PNC) loss. For a given training image containing a lesion and its corresponding mask, we sample representative background patches from the non-lesion regions of the same image. These patches serve as "pseudo-normal" prototypes (samples of normal tissue from an otherwise abnormal image). We then enforce a contrastive objective that maximizes the distance between the embeddings of these background patches and the abnormal lesion features in the embedding space. This patch-level contrastive approach yields more robust lesion localization, as the model learns to distinguish lesion content from normal tissue context within each image, effectively learning a localized normality model from abnormal samples.[25-27]

Endoscopy-specific data augmentation

To address the critical challenge of domain shift in medical imaging, we develop a comprehensive data augmentation strategy designed to simulate the diverse visual conditions encountered in clinical colonoscopy. Our pipeline incorporates a set of transformations that mimic common endoscopic artifacts, thereby enhancing the model's robustness to these variations. These transformations include: (1) Photometric Perturbations: To account for variations in endoscope hardware and lighting, we apply random shifts in white balance and color temperature, as well as gamma correction to model exposure drift. (2) Optical and Motion Artifacts: We simulate specular highlights, a common artifact in mucosal imaging, by adding bright Gaussian blobs to the image. Motion blur is introduced using small directional kernels to reflect artifacts from rapid endoscope or tissue movement. Furthermore, optical effects such as vignetting and mild fisheye distortion are included to model the characteristics of wide-angle endoscope lenses. (3) Compression Artifacts: To mimic the effects of video compression used in clinical archiving systems, we simulate compression-like artifacts by performing a down-sampling and subsequent up-sampling operation. This strategy embeds domain knowledge of the endoscopic procedure into the training process, enhancing the model's robustness to real-world clinical variations. The complete augmentation pipeline with all probabilities and magnitude ranges is provided in Supplementary Table 2.

Results

Training and implementation details

We implemented the proposed method in PyTorch and used a frozen ViT-L/14-336 CLIP backbone with 518 × 518 inputs on a single NVIDIA A5000 GPU.13 Only the lightweight text and visual adapters were trainable. In Stage 1, we trained the text adapter for 20 epochs with Adam (lr = 1×10^{-5}, batch size = 16). The training objective combined segmentation loss (Dice + Focal) with an orthogonality regularizer that separated normal and anomalous embeddings. In Stage 2, we trained the image adapter for 40 epochs with Adam (lr = 5×10^{-4}, batch size = 2) and a multi-step scheduler (milestones at $0.5T$ and $0.75T$). The Stage 2 training objective combined fused segmentation loss, an auxiliary per-level segmentation loss (weight = 0.25), and a PNC loss (weight = 1.0). Complete training hyperparameters and tuning policies are summarized in Supplementary Table 3.

PNC learning: The binary mask was downsampled to the ViT patch grid (37 × 37 tokens for patch size 14). A patch was labeled as *lesion* if its foreground occupancy was ≥ 0.25; pseudo-normal patches were drawn from a 3 × 3 dilated non-lesion ring with occupancy ≤ 0.05. Up to 64 patches of each type were sampled per image. A temperature-scaled ($\tau = 0.07$) patch-to-anchor classification loss was used with a prototype-separation margin of 0.10.

Text prompts: Anchors were built from 5 normal and 5 abnormal prompts combined with 4 templates (20 + 20 text prompts). For colorectal data, the abnormal anchor was enriched with 12 lesion-attribute terms instantiated through 3 templates (36 sentences), fused with the base abnormal anchor at $\alpha = 0.5$. The full prompts are listed in Supplementary Table 1.

Endoscopy-specific augmentation: The pipeline applied horizontal/vertical flips, white-balance jitter ($p = 0.5$), color-temperature shift ($p = 0.5$, ±0.08), gamma correction ($p = 0.5$, 0.8–1.2), synthetic specular highlights ($p = 0.3$; 1–3 Gaussian blobs, radius 3%–8%, intensity 0.10–0.35), motion blur ($p = 0.2$; kernel size 3–7, angle range 0°–180°), vignetting ($p = 0.25$), fisheye distortion ($p = 0.2$), and compression simulation ($p = 0.3$). GeM pooling was initialized at $p = 3.0$ and learned during training; visual gate scalars used sigmoid parameterization initialized at 0.1.

Overall findings

We evaluated the proposed framework under a rigorous leave-one-dataset-out zero-shot protocol. In each run, the model was trained on a single source dataset and evaluated directly, without any target-domain fine-tuning, on the remaining three unseen datasets. We benchmarked our approach against a standard U-Net baseline using pixel-level area under the ROC curve (AUROC) and average precision (AP) as the primary metrics.

As summarized in Table 1, the proposed CLIP-based framework outperformed the U-Net baseline in all 24 AUROC/AP comparisons (12 transfer scenarios × 2 metrics). Averaged over all transfer settings, our method achieved a mean pixel AUROC of 94.34% and a mean pixel AP of 81.14%, compared with 87.22% and 62.21% for U-Net, respectively. The average gains were therefore +7.12 AUROC points and +18.93 AP points. A pooled paired analysis across all leave-one-dataset-out transfer settings further showed mean per-image improvements of +7.41 AUROC points, +19.47 AP points, and +0.231 Dice, with all overall Holm-corrected p-values smaller than 10^{-6}.

Table 1. Leave-one-dataset-out zero-shot colorectal lesion segmentation results at the pixel level

Protocol		Pixel AUROC ↑			Pixel AP (PR) ↑		
Train set	**Zero-shot test set**	**Ours**	**U-Net**	***P* for AUROC**	**Ours**	**U-Net**	***P* for AP**
CVC-ColonDB	CVC-ClinicDB	**96.38**	87.70	$<10^{-6}$	**85.31**	65.32	$<10^{-6}$
	Kvasir-SEG	**94.87**	84.88	$<10^{-6}$	**85.06**	59.50	$<10^{-6}$
	CVC-300	**99.96**	99.11	1.34×10^{-2}	**99.02**	95.55	3.57×10^{-1}
CVC-ClinicDB	CVC-ColonDB	**88.03**	81.77	$<10^{-6}$	**69.50**	52.55	$<10^{-6}$
	Kvasir-SEG	**96.64**	89.31	$<10^{-6}$	**89.76**	69.70	$<10^{-6}$
	CVC-300	**99.75**	95.93	$<10^{-6}$	**92.38**	71.62	4.11×10^{-6}
Kvasir-SEG	CVC-ColonDB	**93.20**	86.34	$<10^{-6}$	**72.20**	54.38	$<10^{-6}$
	CVC-ClinicDB	**96.58**	94.83	$<10^{-6}$	**90.17**	82.12	$<10^{-6}$
	CVC-300	**99.49**	96.64	2.22×10^{-4}	**90.30**	77.03	1.05×10^{-3}
CVC-300	CVC-ColonDB	**85.24**	75.40	$<10^{-6}$	**55.97**	34.31	$<10^{-6}$
	CVC-ClinicDB	**92.36**	76.44	$<10^{-6}$	**72.67**	41.86	$<10^{-6}$
	Kvasir-SEG	**89.39**	78.25	$<10^{-6}$	**71.72**	42.56	$<10^{-6}$

(1) All AUROC and AP values are percentages; ↑ indicates that higher values represent better performance. (2) *P* for AUROC and *P* for AP represent Holm-corrected paired significance values comparing Ours and U-Net at the image/frame level. (3) Extremely small corrected *P*-values are reported as $<10^{-6}$ rather than 0 for readability. (4) Zero-shot cross-dataset protocol: in each run, the model is trained on one source dataset and evaluated directly on the other three target datasets without target-domain fine-tuning. (5) The benchmark datasets are extracted from video sequences; therefore, the reported statistical tests should be interpreted at the image/frame level. AP, average precision; AUROC, area under the receiver operating characteristic curve; PR, precision–recall; U-Net, U-shaped convolutional neural network.

The statistical results further suggest that these improvements are not driven by isolated cases. After Holm correction, the AUROC advantage remained significant in all 12 transfer settings, while the AP advantage remained significant in 11 of 12 settings. The only non-significant AP comparison was CVC-ColonDB → CVC-300 ($P = 3.57 \times 10^{-1}$), where both methods were already close to ceiling performance. Because several benchmarks are frame-based video datasets, these *P*-values should be interpreted as frame-level paired significance rather than patient-level or video-level inference.

Zero-shot cross-dataset performance

We next analyze the transfer behavior by source dataset. Table 1 shows that the strongest average performance was obtained when training on CVC-ColonDB, where our method achieved a mean AUROC of 97.11% and a mean AP of 89.67% across the three unseen targets, compared with 90.56% and 73.46% for U-Net. In this setting, the gains were especially large on CVC-ClinicDB (85.00% vs. 65.32% AP) and Kvasir-SEG (85.13% vs. 59.50% AP). On CVC-300, both methods were close to saturation (98.89% vs. 95.55% AP), which explains why the AP difference in this single transfer was not statistically significant after correction ($P = 3.57 \times 10^{-1}$), despite a still significant AUROC gap ($P = 1.34 \times 10^{-2}$).

When trained on the larger CVC-ClinicDB and Kvasir-SEG source domains, our model also showed stable cross-dataset transfer. For CVC-ClinicDB → Kvasir-SEG, the AP improved from 69.70% to 89.76% (+20.06 points). For Kvasir-SEG as the source, the model maintained AUROCs of 93.20% to 99.49% and APs of 72.20% to 90.30% across all three unseen targets, outperforming U-Net in every case. Importantly, all six AUROC/AP comparisons involving these two source datasets remained significant after Holm correction, with the largest corrected *P*-values still small (2.22×10^{-4} for AUROC and 1.05×10^{-3} for AP).

The most challenging setting remained training on CVC-300. Even in this case, our method achieved a mean AUROC of 89.00% and a mean AP of 66.79%, compared with 76.70% and 39.58% for U-Net, corresponding to improvements of +12.30 AUROC points and +27.21 AP points. The largest AP gains were observed on CVC-ClinicDB (+30.81 points) and Kvasir-SEG (+29.16 points). All corresponding corrected *P*-values were extremely small (e.g., CVC-300 → CVC-ClinicDB: AUROC, $P < 10^{-6}$; AP, $P < 10^{-6}$), supporting that the advantage of the proposed anomaly-aware vision-language model is statistically robust rather than anecdotal. At the same time, because CVC-ClinicDB and CVC-300 are frame-based benchmarks extracted from video sequences, these significance results should be interpreted as frame-level evidence of consistent improvement rather than as direct patient-level or video-level inference.

Comparison with other vision-language backbones

We compare against two additional vision-language backbones, ALBEF and BLIP,[29,30] under the same zero-shot protocol (CVC-ColonDB → CVC-ClinicDB, Kvasir-SEG, CVC-300) with identical text prompts. Two settings are used: **Setting A (basic VL)** trains only a plain segmentation head on the frozen backbone, excluding all proposed modules; **Setting B (backbone replacement)** swaps CLIP for ALBEF/BLIP while retaining our full pipeline. Both use the native 224 × 224 resolution of the public ALBEF/BLIP checkpoints.

As shown in Table 2, our segmentation modules consistently outperform ALBEF/BLIP in their basic-VL experiments (BLIP mean pixel AP rises from 44.07% to 60.36%). Regarding the experimental results, we have the following two observations: (i) Simply replacing the backbone with another generic vision-language encoder does not improve model performance, confirming that the anomaly-aware prompt design, contrastive patch learning, and multi-scale fusion modules are important for performance improvement. (ii) CLIP's contrastive pre-training on large-scale image-text pairs provides a stronger pretrained representation for dense anomaly localization than the masked-language modeling used by ALBEF and BLIP.

Benchmarking against polyp segmentation and medical baselines

We further compared our method with two strong supervised polyp segmentation models, PraNet and TransUNet,[2,31] and two medical vision-language methods, MADCLIP and MedCLIP.[12,32] All methods were trained on CVC-ColonDB and evaluated in a zero-shot setting on CVC-ClinicDB, Kvasir-SEG, and CVC-300 under the same protocol. Since the target splits are positive-only, we report pixel AUROC, pixel AP, and Dice.

Table 3 shows that our method achieves the best mean AUROC and AP among all compared methods. PraNet, the strongest supervised baseline, yields a marginally higher mean Dice (0.8388 vs. 0.8353) but lower mean AUROC (-1.21) and AP (-1.30). Both PraNet and TransUNet remain competitive on the relatively easy CVC-300 transfer yet show a clear gap in the more challenging CVC-ClinicDB and Kvasir-SEG settings, where our method obtains the highest AUROC and AP. MADCLIP and MedCLIP rank below all other methods by a wider margin (MedCLIP mean AP: 59.21). Since they are originally designed for image-level classification rather than dense pixel-level segmentation, their feature embeddings lack the spatial representation needed for precise lesion localization.

Lesion-size detection performance

To evaluate robustness across lesion scales, we evaluated zero-shot results (CVC-ColonDB → {CVC-ClinicDB, Kvasir-SEG, CVC-300}, epoch 40) by lesion-area ratio $r=|\Omega\text{mask}\,|/|\Omega\text{image}\,|$, defining three bins: small ($r < 3\%$), medium ($3\% \le r < 10\%$), and large ($r \ge 10\%$). Since all target sets are positive-only, we report pixel-level AUROC/AP, Dice, and lesion-level recall. A lesion is counted as detected if the binarized anomaly map covers ≥10% of its ground-truth pixels.

Table 4 shows that even for the smallest lesions ($r < 3\%$), recall remains 97.52% on CVC-ClinicDB, 97.40% on Kvasir-SEG, and 100% on CVC-300, with miss rates of at most 2.60%. Segmentation quality is lower for small lesions (Dice 0.59–0.64, AP 62–66 on the two larger datasets), but improves substantially for medium lesions (Dice > 0.83). This suggests that the main failure mode for tiny polyps is inaccurate boundary localization rather than missed detection. Across all three bins and datasets, lesion-level recall stays above 97%, confirming that the framework generalizes well across lesion scales under a strict zero-shot protocol.

Qualitative localization performance

Figure 2 shows representative anomaly heatmaps on three unseen target datasets. The selected examples cover clinically distinct polyp types: small sessile polyps (<5 mm, smooth dome-shaped surface), stalked polyps with irregular heads, and flat elevated lesions (Paris 0-IIa, with subtle color change). Across all three types, the model produces strong activation in the lesion region while suppressing false responses in the surrounding normal tissue. Flat elevated lesions, which are among the most commonly missed findings during colonoscopy, also receive high anomaly scores.

Scoring and thresholding strategy: For threshold-free metrics (pixel AUROC and AP), the raw anomaly maps are normalized across the test set. For Dice computation, each image is independently normalized to [0, 1] via per-image min-max scaling and converted to a binary mask using a fixed threshold of 0.5. This threshold remains constant across all leave-one-dataset-out runs and is not adjusted for any target dataset. In a clinical deployment, threshold calibration on a held-out validation set would be required to balance sensitivity and specificity for the intended operating point.

Engineering efficiency and deployment characteristics

To assess deployment feasibility, we evaluated inference cost for the epoch-40 model in the CVC-ColonDB → CVC-ClinicDB zero-shot setting on a single NVIDIA A5000 GPU at 518 × 518 resolution. Latency was measured with batch size 1 over 30 timed iterations (after 10 warm-up runs), excluding data-loading overhead. End-to-end throughput was additionally measured on the full CVC-ClinicDB test set with batch size 8.

As Table 5 shows, only 2.85% of parameters (12.58M) are introduced by the proposed adapters, fusion module, and GeM pooling; the frozen ViT-L/14 backbone dominates computational cost. The deployed checkpoint is approximately 1 GB with 2 GB peak GPU memory usage. Inference latency is approximately 113 ms per image (~8.9 img/s). At the current stage, the system serves as a research prototype and does not yet meet the 25–30 fps requirement for real-time clinical use. Since the frozen backbone accounts for over 97% of the total computation, reducing inference time should focus on backbone-level optimization (e.g., lighter vision-language encoders, mixed-precision inference, and temporal feature reuse across video frames) rather than removing the lightweight task-specific modules.

Ablation study

We ablated three core components: (i) removing the PNC loss, (ii) replacing attribute-enhanced prompts with generic prompts, and (iii) disabling the enhanced visual branch by reverting to fixed gates, mean fusion, and average pooling. All variants were trained on CVC-ColonDB and evaluated in a zero-shot setting on CVC-ClinicDB, Kvasir-SEG, and CVC-300 using the same protocol and hyperparameters as the full model.

Table 6 shows that the full model achieves the highest mean Dice (0.8329). Removing attribute-enhanced prompts leads to a small but consistent drop in AP (89.87 → 89.52) and Dice (0.8329 → 0.8322), confirming that domain-specific textual descriptions provide complementary semantic guidance beyond generic prompts. Removing the PNC loss slightly increases AP on Kvasir-SEG but lowers CVC-ClinicDB AUROC (96.28 → 94.81) and CVC-300 Dice (0.9112 → 0.8947), resulting in lower mean AUROC and Dice overall. The contrastive objective thus

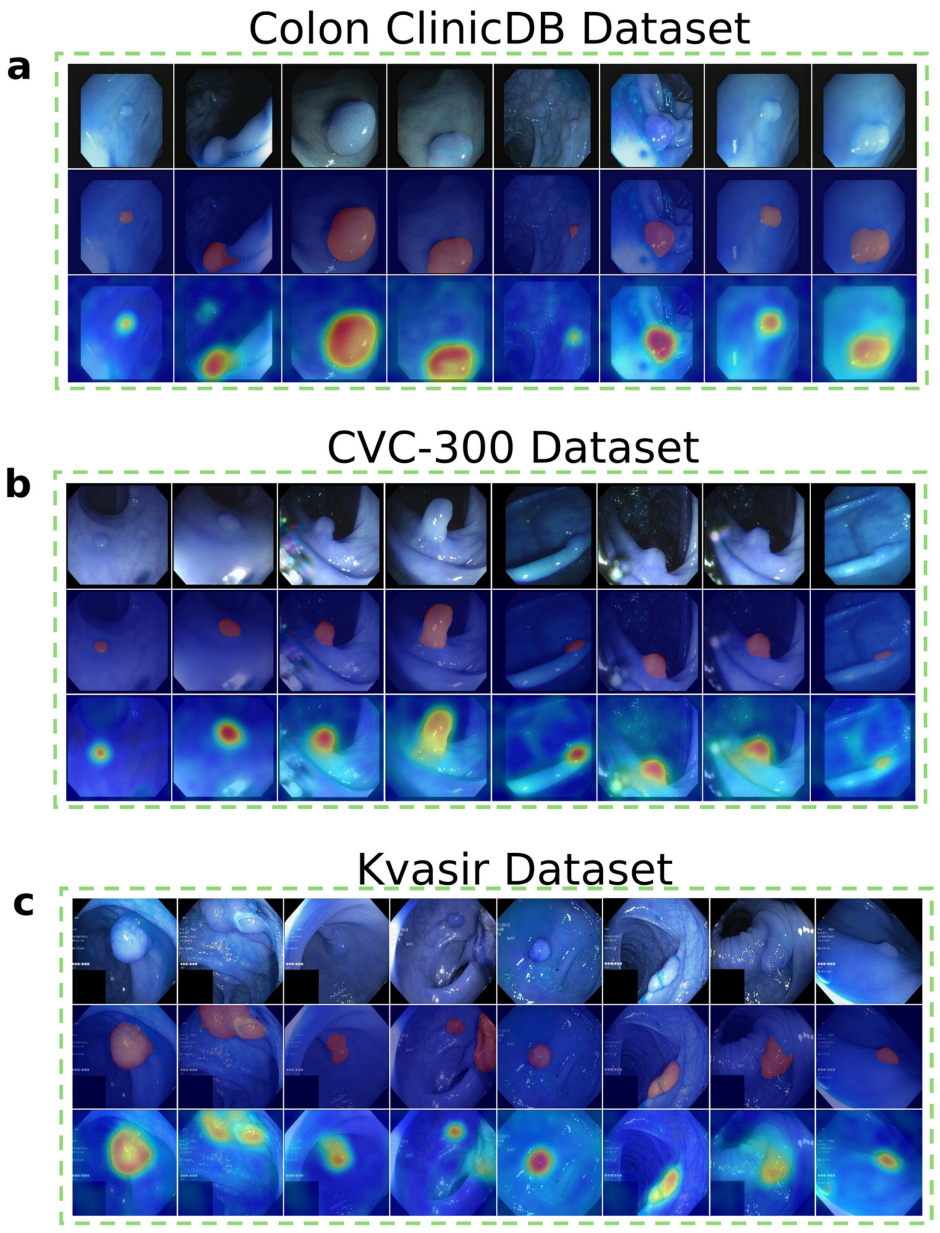

Fig. 2. Qualitative zero-shot lesion localization on three unseen target datasets: (a) CVC-ClinicDB, (b) CVC-300, and (c) Kvasir-SEG. Top row: input images; middle row: ground-truth masks (red); bottom row: predicted anomaly heatmaps (red/yellow: high anomaly score, blue: normal background). The examples span a range of clinically relevant morphologies, including small sessile polyps with smooth dome-shaped surfaces, pedunculated polyps attached by a stalk, and flat elevated (0-IIa) lesions with only subtle mucosal color change. Despite variations in lesion size, shape, lighting, and mucosal texture, the model consistently localizes lesion regions and suppresses false activations on normal tissue, all without target-domain fine-tuning.

primarily improves robustness under cross-dataset shift. Removing the enhanced visual branch produces the lowest mean Dice (0.8268) and the largest localization drop on CVC-300 (Dice 0.9112 → 0.8942), while AUROC changes minimally, indicating that adaptive fusion and GeM pooling mainly benefit spatial localization rather than coarse anomaly ranking. Each component addresses a different aspect of cross-dataset generalization, and the full combination yields the best balanced performance.

Discussion

Future directions

This work establishes a strong foundation for the development of generalizable AI in colonoscopy. A logical next step is to extend this validation to real-time video sequences. Evaluating the performance of the model on full-length colonoscopy videos will be crucial for assessing its temporal consistency, processing speed,

Table 2. Zero-shot comparison with ALBEF and BLIP backbones (trained on CVC-ColonDB). Each entry: Pixel AUROC/Pixel AP/Dice

Method	CVC-ClinicDB	Kvasir-SEG	CVC-300	Average
ALBEF-basic VL	79.27/32.81/0.467	80.79/52.53/0.530	97.71/62.23/0.396	85.92/49.19/0.464
BLIP-basic VL	81.34/41.29/0.421	76.37/46.76/0.480	93.23/44.16/0.259	83.65/44.07/0.387
ALBEF+Ours	81.50/42.08/0.452	76.32/49.63/0.466	98.51/70.85/0.405	85.44/54.19/0.441
BLIP+Ours	86.46/53.42/0.445	80.41/53.37/0.516	98.19/74.30/0.339	88.35/60.36/0.433
Ours (CLIP)	**96.38/85.31/0.787**	**94.87/85.06/0.808**	**99.96/99.02/0.911**	**97.07/89.80/0.835**

ALBEF, Align Before Fuse; AP, average precision; AUROC, area under the receiver operating characteristic curve; BLIP, Bootstrapping Language–Image Pre-training; CLIP, Contrastive Language–Image Pre-training; Dice, Dice similarity coefficient; VL, vision–language.

Table 4. Lesion-size stratified zero-shot performance (trained on CVC-ColonDB). A lesion is detected when the thresholded anomaly map covers ≥10% of the ground-truth mask

Test set	Size bin	Ratio range	N	Mean ratio (%)	Pixel AUROC	Pixel AP	Dice	Recall	Miss rate
CVC-ClinicDB	Bin-1	[0.00%, 3.00%)	121	1.94	97.51	66.41	0.6360	97.52	2.48
CVC-ClinicDB	Bin-2	[3.00%, 10.00%)	283	5.97	98.47	89.58	0.8384	98.59	1.41
CVC-ClinicDB	Bin-3	[10.00%, 100.00%]	208	17.71	94.95	87.44	0.7942	99.52	0.48
Kvasir-SEG	Bin-1	[0.00%, 3.00%)	77	2.10	97.73	62.12	0.5853	97.40	2.60
Kvasir-SEG	Bin-2	[3.00%, 10.00%)	355	6.22	98.78	90.15	0.8300	99.72	0.28
Kvasir-SEG	Bin-3	[10.00%, 100.00%]	560	22.98	94.21	87.49	0.8178	99.64	0.36
CVC-300	Bin-1	[0.00%, 3.00%)	35	2.09	99.96	98.34	0.8828	100.00	0.00
CVC-300	Bin-2	[3.00%, 10.00%)	24	4.60	99.96	99.30	0.9501	100.00	0.00
CVC-300	Bin-3	[10.00%, 100.00%]	1	18.45	99.95	99.77	0.9686	100.00	0.00

CVC-300 contains only one large-lesion case; this subgroup should be interpreted cautiously. AP, average precision; AUROC, area under the receiver operating characteristic curve; Dice, Dice similarity coefficient.

Table 3. Zero-shot benchmarking (trained on CVC-ColonDB). Mean is computed over the three target datasets

Method	Mean AUROC	Mean AP	Mean Dice
Ours	**97.07**	**89.80**	0.8353
PraNet	95.86	88.50	**0.8388**
TransUNet	94.95	87.57	0.8369
MADCLIP	94.75	84.69	0.7452
MedCLIP	90.73	59.21	0.5151

AP, average precision; AUROC, area under the receiver operating characteristic curve; CLIP, Contrastive Language–Image Pre-training; Dice, Dice similarity coefficient.

and practical utility in a clinical workflow. More broadly, the intelligent oncology literature is increasingly integrating computational pathology and radiomics-based response modeling into precision cancer workflows, including recent work on digital and intelligent pathology and neoadjuvant-response prediction in rectal cancer.[33,34] Furthermore, the vision-language framework of the model opens promising avenues for more advanced diagnostic tasks. Future research will explore adapting the model to perform zero-shot classification of detected lesions into clinically relevant subtypes, such as adenomatous versus hyperplastic lesions. Such a capability would provide endoscopists with real-time decision support, enabling the adoption of "resect and discard" protocols for low-risk lesions, which could improve procedural efficiency and reduce pathology costs.[11,12,14,23]

Limitations

Several limitations should be noted. First, all four benchmark datasets contain only colorectal polyps. Publicly available endoscopy segmentation benchmarks with pixel-level masks for non-polyp colorectal pathologies—such as ulcerative colitis, colorectal tumors, or inflammatory erosions—do not currently exist. Datasets like HyperKvasir provide classification labels for ulcerative colitis (Mayo grades) and colorectal cancer but do not include segmentation masks for these categories. The scope of the present evaluation is therefore limited to polyp detection and localization; whether the anomaly-aware design generalizes to other mucosal abnormalities remains an open question that will require new annotated data, ideally produced through multicenter clinical collaborations. Second, all four datasets originate from European academic centers, introducing potential population and hardware bias. Third, the evaluation uses curated still frames rather than full-procedure video, which does not capture motion artifacts, transient views, or the predominance of polyp-free frames in real colonoscopies. Fourth, the current inference speed (≈8.9 fps) falls short of the 25–30 fps needed for real-time overlay;

Table 5. Engineering characteristics of the proposed model (epoch 40, CVC-ColonDB → CVC-ClinicDB, single A5000 GPU, 518 × 518 input)

Metric	Value
Total/added parameters	441.34 M/12.58 M (2.85%)
Deployed checkpoint size	≈1007.84 MB
Peak CUDA memory	2009.42 MB
FLOPs/MACs per image	1041.73 G/520.86 G
Mean latency (p50/p90)	112.90 ms (112.91/113.25)
Single-image throughput	8.86 img/s
End-to-end throughput (CVC-ClinicDB)	8.94 img/s

CUDA, Compute Unified Device Architecture; FLOPs, floating-point operations; GPU, graphics processing unit; MACs, multiply-accumulate operations; p50/p90, 50th and 90th percentile latency, respectively.

Table 6. Ablation study on CVC-ColonDB → {CVC-ClinicDB, Kvasir-SEG, CVC-300}. Values are averaged over the three unseen target sets

Variant	Pixel AUROC ↑	Pixel AP ↑	Dice ↑
Full model	97.07	89.80	**0.8353**
w/o PNC loss	96.72	**90.05**	0.8313
w/o attribute prompts	97.07	89.52	0.8322
w/o visual enhancement	**97.12**	89.81	0.8268

↑ indicates that higher values represent better performance. AP, average precision; AUROC, area under the receiver operating characteristic curve; Dice, Dice similarity coefficient; PNC loss, pseudo-normal contrastive loss; w/o, without.

the bottleneck is the frozen ViT-L/14 backbone. Fifth, there is no same-dataset blinded reader comparison with human endoscopists. Finally, the anomaly-detection formulation treats all abnormal tissue as a single class and does not distinguish adenomas from hyperplastic polyps. Taken together, these factors mean that the results demonstrate zero-shot technical generalization across polyp datasets but do not yet constitute clinical validation.

Clinical workflow integration and regulatory considerations

The intended translational role of the proposed method is adjunctive CADe rather than autonomous diagnosis. In practice, the model would process the live white-light video stream frame by frame and overlay a heatmap on suspicious regions, while the endoscopist retains full responsibility for interpretation, biopsy, and surveillance decisions. This is consistent with existing CADe systems designed to support rather than replace clinical judgment.[5,9] Clinical deployment would further require threshold calibration and temporal smoothing to suppress false positives across consecutive frames, as well as prospective reader studies to evaluate whether the model reduces miss rates among endoscopists at varying experience levels.

From a regulatory standpoint, the framework falls under Software as a Medical Device (SaMD). International SaMD guidance requires evidence of valid clinical association, analytical validation, and clinical validation in the intended-use population. For a colonoscopy CADe tool, this typically involves multicenter testing under diverse imaging conditions, human factors assessment, and post-deployment monitoring for performance drift. The present work is a retrospective still-image study without a same-dataset reader comparison against human experts; the results should therefore be interpreted as evidence of technical zero-shot generalization rather than as demonstrating superiority over endoscopists.

Conclusions

This study demonstrates that an anomaly-aware vision-language framework significantly improves zero-shot generalization for colorectal lesion detection without requiring dataset-specific fine-tuning or normal training images. Without any dataset-specific fine-tuning, the model achieved a mean pixel-level AUROC of 0.94 and a mean average precision of 0.81 across the 12 leave-one-dataset-out zero-shot transfer settings. In the CVC-ColonDB-source benchmark, the model achieved a mean Dice coefficient of 0.84 across CVC-ClinicDB, Kvasir-SEG, and CVC-300. These results underscore the model's ability to accurately localize diverse polyp morphologies across heterogeneous clinical scenarios that it was not explicitly trained to recognize. Although further work is needed to address technical limitations such as real-time inference and inclusion of non-polyp pathologies, this multimodal strategy establishes a solid technical foundation for the deployment of robust AI-based colorectal screening systems.

Supporting information

Supplementary material for this article is available at https://doi.org/10.14218/OnA.2026.00006.

Acknowledgments

None.

Funding

This work was supported by the National Natural Science Foundation of China (82405031 to SCT).

Conflict of interest

Dr. Sichao Tian serves as an editorial board member of *Oncology Advances*. The authors declare that they have no other conflicts of interest related to this publication.

Author contributions

Conceptualization, methodology, software implementation, writing of the original draft (ZH, XC), investigation, data curation, formal analysis, validation of the experimental results (MM, BL), overall supervision, review and editing of the manuscript (WZ, JZ, ST). All authors have approved the final version and publication of the manuscript.

Ethical statement

All data used in this study were obtained from publicly available, de-identified datasets. No identifiable personal information was accessed in this study. Therefore, this study did not require additional ethical approval or informed consent. All procedures performed in this study were in accordance with the principles of the 2024 Declaration of Helsinki.

Data sharing statement

The datasets used to support the findings of this study are available in the following repositories: CVC-ClinicDB repository (available at https://www.kaggle.com/datasets/balraj98/cvcclinicdb); CVC-ColonDB repository (DOI: 10.1109/TMI.2014.2314959); Kvasir-SEG repository (available at https://datasets.simula.no/kvasir-seg/); and CVC-300 repository (DOI: 10.1155/2017/4037190). The technical code and models used to support the findings of this study are available from the corresponding author upon request.

References

[1] Zhao S, Wang S, Pan P, Xia T, Chang X, Yang X, *et al*. Magnitude, Risk Factors, and Factors Associated with Adenoma Miss Rate of Tandem Colonoscopy: A Systematic Review and Meta-analysis. Gastroenterology 2019;156(6):1661–1674.e11. DOI: 10.1053/j.gastro.2019.01.260, PMID: 30738046.

[2] Fan DP, Ji GP, Zhou T, Chen G, Fu H, Shen J, *et al*. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, *et al*., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020 Lecture Notes in Computer Science, vol 12266. Cham: Springer; 2020. p. 263-273. DOI: 10.1007/978-3-030-59725-2_26.

[3] Ali S, Jha D, Ghatwary N, Realdon S, Cannizzaro R, Salem OE, *et al*. A multi-centre polyp detection and segmentation dataset for generalisability assessment. Sci Data 2023;10(1):75. DOI: 10.1038/s41597-023-01981-y, PMID: 36746950.

[4] Ali S, Ghatwary N, Jha D, Isik-Polat E, Polat G, Yang C, *et al*. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. Sci Rep 2024;14(1):2032. DOI: 10.1038/s41598-024-52063-x, PMID: 38263232.

[5] Jiang J, Xie Q, Cheng Z, Cai J, Xia T, Yang H, *et al*. AI based colorectal disease detection using real-time screening colonoscopy. Precis Clin Med 2021;4(2):109–118. DOI: 10.1093/pcmedi/pbab013, PMID: 35694157.

[6] Zhao K, Wu L, Huang Y, Yao S, Xu Z, Lin H, *et al*. Deep learning quantified mucus-tumor ratio predicting survival of patients with colorectal cancer using whole-slide images. Precis Clin Med 2021;4(1):17–24. DOI: 10.1093/pcmedi/pbab002, PMID: 35693123.

[7] Wang H, Zhou Z, Li H, Xiang W, Lan Y, Dou X, *et al*. Blood Biomarkers Panels for Screening of Colorectal Cancer and Adenoma on a Machine Learning-Assisted Detection Platform. Cancer Control 2023;30:10732748231222109. DOI: 10.1177/10732748231222109, PMID: 38146088.

[8] Li R, Hao X, Diao Y, Yang L, Liu J. Explainable Machine Learning Models for Colorectal Cancer Prediction Using Clinical Laboratory Data. Cancer Control 2025;32:10732748251336417. DOI: 10.1177/10732748251336417, PMID: 40334702.

[9] Shakir T, Kader R, Bhan C, Chand M. AI in colonoscopy - detection and characterisation of malignant polyps. Art Int Surg 2023;3:186–194. DOI: 10.20517/ais.2023.17.

[10] Ferrari S, Negro S, Celotto F, Bao QR, Madeo G, Pulvirenti A, *et al*. Artificial intelligence for post-polypectomy surveillance: a scoping review of emerging tools in colorectal cancer prevention. Art Int Surg 2025;5:490–504. DOI: 10.20517/ais.2025.65.

[11] Jang J, Kyung D, Kim SH, Lee H, Bae K, Choi E. Significantly improving zero-shot X-ray pathology classification via fine-tuning pre-trained image-text encoders. Sci Rep 2024;14(1):23199. DOI: 10.1038/s41598-024-73695-z, PMID: 39369048.

[12] Shiri M, Beyan C, Murino V. MadCLIP: Few-shot medical anomaly detection with CLIP. In: Gee JC, Alexander DC, Hong J, Iglesias JE, Sudre CH, Venkataraman A, *et al*., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2025 Cham: Springer; 2026. Lecture Notes in Computer Science; vol 15965. p. 416-426. DOI: 10.1007/978-3-032-04978-0_40.

[13] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, *et al*. Learning transferable visual models from natural language supervision. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning (ICML), vol 139. PMLR; 2021. p. 8748-8763.

[14] Huang SC, Shen L, Lungren MP, Yeung S. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal (QC), Canada; 2021 Oct 10-17. IEEE; 2021. p. 3922-3931. DOI: 10.1109/ICCV48922.2021.00391.

[15] Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng 2022;6(12):1399–1406. DOI: 10.1038/s41551-022-00936-9, PMID: 36109605.

[16] Turshudzhyan A, Gornick D, Mertz G, Tadros M. Current Practice and Emerging Endoscopic Technology in the Diagnosis of Colorectal Cancer: A Narrative Review of Enhanced Imaging and Optical Biopsy. J Transl Gastroenterol 2024;2(3):150–158. DOI: 10.14218/JTG.2024.

00011.

[17] Weng E, Dharan M. Endoscopic Resection of Gastrointestinal Lesions: Preference and Feasibility of En bloc Resection Techniques. J Transl Gastroenterol 2023;1(1):40–46. DOI: 10.14218/JTG.2023.00001.

[18] Ma W, Zhang X, Yao Q, Tang F, Wu C, Li Y, *et al*. AA-CLIP: Enhancing zero-shot anomaly detection via anomaly-aware CLIP. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025; 2025 Jun 10-17; Nashville, TN, USA. Piscataway (NJ):IEEE; 2025. p. 4744–4754. DOI: 10.1109/CVPR52734.2025.00447.

[19] Gao P, Geng S, Zhang R, Ma T, Fang R, Zhang Y, *et al*. CLIP-Adapter: Better vision-language models with feature adapters. Int J Comput Vis 2024;132(2):581–595. DOI: 10.1007/s11263-023-01891-x.

[20] Zhou K, Yang J, Loy CC, Liu Z. Learning to prompt for vision-language models. Int J Comput Vis 2022;130(9):2337–2348. DOI: 10.1007/s11263-022-01653-1.

[21] Rao Y, Zhao W, Chen G, Tang Y, Zhu Z, Huang G, *et al*. DenseCLIP: Language-Guided Dense Prediction With Context-Aware Prompting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 19-24; New Orleans, LA, USA. New York: IEEE; 2022. p. 18082-18091.

[22] Zhang R, Zhang W, Fang R, Gao P, Li K, Dai J, *et al*. Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. Computer Vision – ECCV 2022 Lecture Notes in Computer Science, vol 13695. Cham: Springer; 2022. p. 493-510. DOI: 10.1007/978-3-031-19833-5_29.

[23] Li LH, Zhang P, Zhang H, Yang J, Li C, Zhong Y, *et al*. Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 19-24; New Orleans, LA, USA. New York: IEEE; 2022. p. 10965-10975.

[24] Liang W, Zhang Y, Kwon Y, Yeung S, Zou J. Mind the gap: understanding the modality gap in multi-modal contrastive representation learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022); 2022 Nov 28-Dec 9; New Orleans, LA, USA. Red Hook (NY): Curran Associates, Inc.; 2022. p. 17612-17625.

[25] Roth K, Pemula L, Zepeda J, Schölkopf B, Brox T, Gehler P. Towards Total Recall in Industrial Anomaly Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 19-24; New Orleans, LA, USA. New York: IEEE; 2022. p. 14298-14308. DOI: 10.1109/CVPR52688.2022.01392.

[26] Zavrtanik V, Kristan M, Skočaj D. DRAEM – a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 11-17; Los Alamitos, CA, USA. New York: IEEE Computer Society; 2021. p. 8310-8319. DOI: 10.1109/ICCV48922.2021.00822.

[27] Bergmann P, Fauser M, Sattlegger D, Steger C. MVTec AD: A comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 16-20; Long Beach, CA, USA. New York: IEEE; 2019. p. 9592-9600.

[28] Cheng Z, Li Y, Chen H, Zhang Z, Pan P, Cheng L. DSGMFFN: Deepest semantically guided multi-scale feature fusion network for automated lesion segmentation in ABUS images. Comput Methods Programs Biomed 2022;221:106891. DOI: 10.1016/j.cmpb.2022.106891, PMID: 35623209.

[29] Li J, Selvaraju RR, Gotmare AD, Joty S, Xiong C, Hoi SCH. Align before fuse: vision and language representation learning with momentum distillation. In: Advances in Neural Information Processing Systems (NeurIPS 2021); 2021 Dec 6-14. Red Hook (NY): Curran Associates Inc; 2021. p. 9694-9705.

[30] Li J, Li D, Xiong C, Hoi SCH. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the 39th International Conference on Machine Learning (ICML); 2022 Jul 17-23; Baltimore, MD, USA. Proceedings of Machine Learning Research, vol 162. p. 12888-12900.

[31] Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, *et al*. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers Med Image Anal. 2024; 97:103280. DOI: 10.1016/j.media.2024.103280.

[32] Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. Proc Conf Empir Methods Nat Lang Process 2022;2022:3876–3887. DOI: 10.18653/v1/2022.emnlp-main.256, PMID: 39144675.

[33] Huang Q, Wu S, Ou Z, Gao Y. Computational pathology: a comprehensive review of recent developments in digital and intelligent pathology. Intell Oncol 2025;1(2):139–159. DOI: 10.1016/j.intonc.2025.03.004.

[34] Liu B, Feng J, Hu Y, Tang R, Zhang Y, Wang Y, *et al*. Predicting the effectiveness of neoadjuvant therapy in rectal cancer patients: model construction based on radiomics and carcinoembryonic antigens. Intell Oncol 2026;2(1):100035. DOI: 10.1016/j.intonc.2025.12.003.