Colon-Bench: An Agentic Workflow for Scalable Dense Lesion Annotation in Full-Procedure Colonoscopy Videos

Abdullah Hamdi, Changchun Yang, and Xin Gao

King Abdullah University of Science and Technology (KAUST)
{abdullah.hamdi@kaust.edu.sa}

Abstract. Early screening via colonoscopy is critical for colon cancer prevention, yet developing robust AI systems for this domain is hindered by the lack of densely annotated, long-sequence video datasets. Existing datasets predominantly focus on single-class polyp detection and lack the rich spatial, temporal, and linguistic annotations required to evaluate modern Multimodal Large Language Models (MLLMs). To address this critical gap, we introduce Colon-Bench, generated via a novel multi-stage agentic workflow. Our pipeline seamlessly integrates temporal proposals, bounding-box tracking, AI-driven visual confirmation, and human-in-the-loop review to scalably annotate full-procedure videos. The resulting verified benchmark is unprecedented in scope, encompassing 528 videos, 14 distinct lesion categories (including polyps, ulcers, and bleeding), over 300,000 bounding boxes, 213,000 segmentation masks, and 133,000 words of clinical descriptions. We utilize Colon-Bench to rigorously evaluate state-of-the-art MLLMs across lesion classification, Open-Vocabulary Video Object Segmentation (OV-VOS), and video Visual Question Answering (VQA). The MLLM results demonstrate surprisingly high localization performance in medical domains compared to SAM-3. Finally, we analyze common VQA errors from MLLMs to introduce a novel "colon-skill" prompting strategy, improving zero-shot MLLM performance by up to 9.7% across most MLLMs. The dataset and the code are available at https://abdullahamdi.com/colon-bench

Keywords: colonoscopy · lesion detection · polyps · MLLMs.

1 Introduction

Colon cancer is the second leading cause of cancer death worldwide [24], and many of these deaths are preventable with early screening and removal of precancerous lesions; however, colonoscopy remains expensive, invasive, and logistically difficult to arrange, requiring specialized clinicians and often long procedures (up to two hours) that increase facility and staffing costs. Automated AI analysis promises to ease these burdens, but lesions are sparse and frequently obscured by motion blur, occlusion, debris, stool, or fluids, and by camera contact with the colon wall,

Table 1. Comparison of Colonoscopy Datasets. Relative to prior datasets that emphasize single-class polyp detection or anatomical segmentation, *Colon-Bench* provides a broader lesion taxonomy and richer supervision. It combines long-sequence coverage with dense boxes, masks, and clinical text, enabling multi-task evaluation across detection, video segmentation, and language-based understanding.

Attribute	Kvasir-SEG [14]	SUN [20]	PolypGen [1]	REAL-Col. [3]	CAS-Col. [23]	Colon-Bench (Ours)
Year	2020	2021	2023	2024	2025	2026
Focus	Segmentation	Detection	Segmentation	Detection	Anatomy	**Lesion, Video Seg., & Text**
Number of Videos	N/A	N/A	N/A	60 (Full-Procedure)	78 (Clip)	**528 (Clip)**
Total Frames/Images	1,000	158,690	6,282	**2,757,723**	1,961,100	464,035
Lesion Classes	1 (Polyp)	1 (Polyp)	1 (Polyp)	1 (Polyp)	N/A	**14 (Bleeding, Polyps, ...)**
Bounding Boxes	N/A	158,690	6,282	**351,264**	N/A	300,132
Segmentation Masks	1,000	N/A	6,282	N/A	Yes	**213,067**
Language Descriptions	N/A	N/A	N/A	N/A	N/A	**Yes (133k Words)**

making dense visual analysis challenging. Moreover, manual dense annotation of colonoscopy videos is labor-intensive and inconsistent, which motivates our work on a scalable, affordable pipeline for dense colonoscopy video annotation, facilitating AI research, evaluation, and applications in colonoscopy.

To address the visual challenges of colonoscopy videos, recent methods specialize architectures for pixel-level precision and subtle lesion identification [28,17]. Because procedures are long, efficient spatiotemporal analysis is also critical; state-space models [25] and token merging [27] have been introduced to capture long-range dependencies while reducing redundancy. Many methods rely on self-supervised pretraining [15] or use limited annotated examples with text [13], bounding boxes [7,16], or segmentation masks [14,7]. These workarounds highlight the gap our work addresses: the need for a comprehensive, densely annotated, long-sequence dataset to ground spatiotemporal analysis of real colonoscopies.

REAL-COLON [3] introduced 60 long colonoscopy sequences labeled with 351,264 polyp bounding boxes, while CAS-COLON [23] provides anatomical temporal classification over 78 videos, 9 hours, and around 2 million images. Other datasets are either too small for comprehensive training and evaluation [19] or large but limited in task scope and lesion diversity [14,20,19,1]. EndoBench [18], a contemporary related work, evaluates *static-image* VQA in general endoscopy. As Table 1 shows, Colon-Bench provides dense lesion annotations with text descriptions and segmentation mask tracklets across diverse video clips, enabling benchmarks for Video Question Answering, lesion clip classification, and Open-Vocabulary Video Segmentation in colonoscopy videos.

Large Language Models (LLMs) [21,11] and foundational vision-language models [22] have accelerated Multimodal Large Language Models (MLLMs) for complex visual reasoning [2,8,10,12,11]. Despite strong zero-shot performance in general domains, their ability to interpret long, occluded, noisy medical sequences remains largely untested, motivating *Colon-Bench* as a systematic benchmark for colonoscopy video understanding. Recent work on agentic MLLM pipelines for long-form videos [9,26] and hybrid automatic labeling with human verification for large medical image datasets [6] further motivates our pipeline in Fig. 1.

Contributions: (i) We introduce a novel, multi-stage agentic workflow for dense annotation of full-procedure colonoscopy videos. By combining temporal

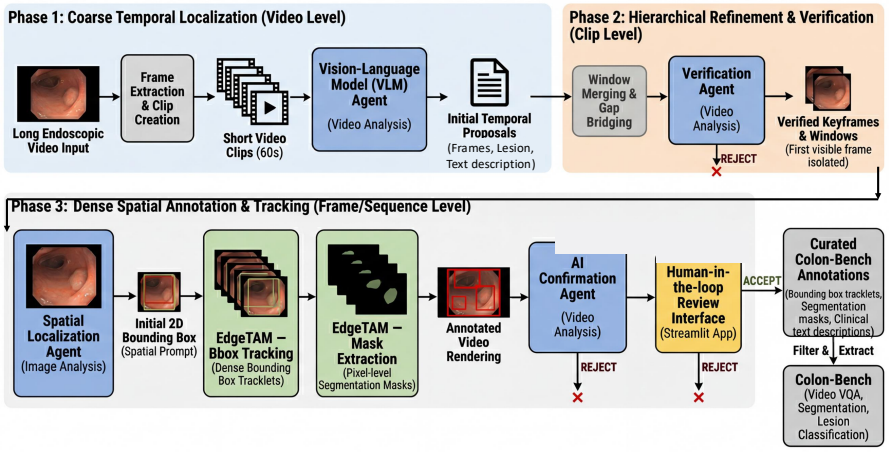

Fig. 1. Colon-Bench Annotation Pipeline. Overview of the multi-stage agentic workflow used to build the dataset, from VLM proposal generation through verification, spatial tracking, AI confirmation, and clinician review. The accepted annotations support multiple video tasks: Visual Quesion Answering (VQA), binary lesion classification, and Open-Vocabulary Video Object Segmentation (OV-VOS).

proposals, bounding-box tracking, AI-driven confirmation with visual cues, and an efficient human-in-the-loop review interface, our pipeline reduces manual labor while yielding high-quality spatial, temporal, and textual annotations. **(ii)** We present a comprehensive, multi-task video benchmark (*Colon-Bench*) for colonoscopy understanding. Spanning 14 distinct lesion categories (including polyps, bleeding, and ulcers), it includes over 464k frames, 300k bounding boxes, 213k segmentation masks, and 133k words of clinical descriptions, all verified by humans and an experienced surgeon. It supports four rigorous evaluation tasks: binary lesion classification, Open-Vocabulary Video Object Segmentation (OV-VOS), and two difficulty tiers of Video Visual Question Answering (VQA). **(iii)** We conduct a large-scale evaluation of lesion-focused colonoscopy visual reasoning and localization in frontier MLLMs. By analyzing cross-model error patterns, we propose a novel "colon-skill" that defeines a textual guidance prompt, improving most MLLM performance by up to 9.7% without additional training.

2 Methodology

2.1 Agentic Workflow for Colonoscopy Annotation

Automatic annotations with quality control. As illustrated in Fig. 1, we built dense colonoscopy annotations by applying a multi-stage agentic pipeline to 60 REAL-COLON videos [3]. An initial vision-language model proposed 1,325 candidate lesion windows (22.97 hours), which were progressively refined by verification filtering, EdgeTAM [29] bounding-box tracking, cued AI confirmation using

Table 2. Annotation Pipeline Stages. We show Colon-Bench annotations after each major processing stage. As automated filters and human review progressively discard false-positive windows (reducing total frames), the temporal precision, F1 score, and specificity of the automatically identified lesions demonstrably increase on the surrogate polyp labels of REAL-COLON [3] (polyps are one type of lesion in Colon-Bench).

Metric	Initial VLM Proposals	+ verification Agent Filtering	+ Cued AI Confirmation Agent	Final Human-Curated Annots.
Total Windows	1,325	903	597	**528**
Total Frames	826,763	648,440	492,606	**464,035**
Duration (Hours)	22.97	18.01	13.68	**12.89**
Total Bounding Boxes	-	-	314,408	**300,132**
Total Seg. Masks	-	-	227,343	**213,067**
Total Clinical Words	-	-	145,515	**133,289**
Precision	30.9	36.2	53.1	**55.4**
Recall	67.5	68.7	48.8	**48.6**
F1 Score	42.4	47.4	50.8	**51.8**
Specificity	78.6	82.9	93.9	**94.4**

lesion-box overlays, and final human review (Table 2). Tracking and AI confirmation produced the spatial annotations (over 314k initial bounding boxes) while pruning weak temporal boundaries. Stage-wise evaluation on REAL-COLON surrogate polyp labels shows that the automated AI stages provide the main precision, F1, and specificity gains; because many Colon-Bench categories (ulcers, bleeding, angiectasia, etc.) are absent from REAL-COLON, these labels serve as quality control rather than exhaustive recall targets. Motivated by Gemini-3's recent medical-domain performance [4], we used Gemini-2.5-flash-lite for proposals, Gemini-3-pro for video verification, and Gemini-3-flash for bounding boxes, balancing performance and cost (Table 3).

Human review. Human review served as the final quality gate. To enable efficient review at scale, the pipeline pre-rendered short clips with spatial overlays into an interactive web interface. A reviewing physician rejected only 69 of the 597 presented windows (11.6%), demonstrating strong agreement with the AI filters. Ultimately, the pipeline retained 528 curated windows (39.8% retention) spanning 464,035 frames. The final dataset yields a dense, multi-modal benchmark comprising 300,132 bounding boxes, 213,067 segmentation masks, and over 133k words of verified clinical text (averaging 252.4 words per window).

Lesion types. The dataset spans 14 lesion categories identified through multi-label keyword matching over clinician-verified text fields (Fig. 2). The distribution is heavily long-tailed: sessile polyps dominate (411 windows), followed by bleeding (252), ulcers (160), and erythematous lesions (112). Because Colon-becnh descriptions are free-form text (one window can have multiple lesions), a single review window may contribute to more than one category.

2.2 Colon-Bench Evaluation Benchmark

Filtering. Colon-Bench is a multi-task video benchmark for colonoscopy understanding comprising 1,597 clips from 60 patients (955,126 frames), spanning

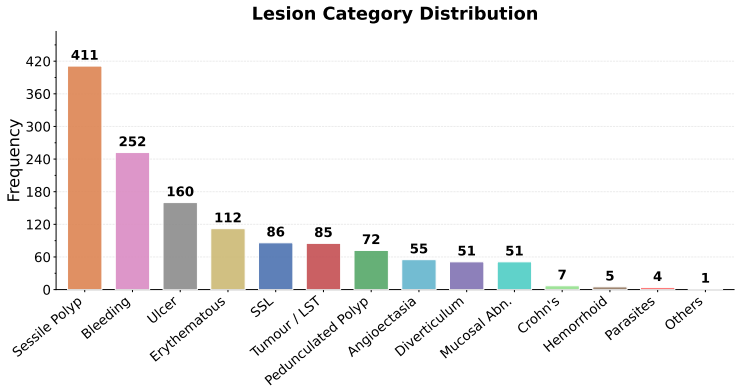

Fig. 2. Lesion Category Distribution. Long-tailed lesion category distribution in Colon-Bench, highlighting the diversity of lesions.

five tasks: binary classification (790 clips), detection (272 clips, 61,538 per-frame bounding boxes), instance segmentation (264 clips, 57,550 per-frame masks), and VQA at two difficulty tiers. The binary lesion classes and segmentation masks follow directly from Sec. 2.1 to establish the first colonoscopy Open-Vocabulary Video Object Segmentation (OV-VOS) benchmark.

VQA (prompted & unprompted) . The *prompted* VQA split contains 1,485 five-choice questions over 499 clips whose videos include bounding-box overlays on confirmed lesion windows; the *unprompted* split contains 2,740 questions over 918 clips rendered from raw frames, fully encompassing all detection and segmentation videos. Across both splits the questions are dominated by lesion type (39%/43%, prompted/unprompted), color & surface appearance (38%/44%), and anatomical location (30%/35%), followed by morphology (20%/20%), with explicit temporal references largely confined to the prompted split (59% vs. 2%). For instance, an unprompted question for the clip in Fig. 3 first row asks: *"How is the morphology and anatomical location of the identified lesion described?"* with distractors spanning pedunculated, depressed, and flat lesions at various anatomical sites; the correct answer is *(D) a sessile polyp on a haustral fold.*

Debiasing Colon-Bench. For each clip, three five-way MCQs are generated with Gemini-3, covering lesion identification, clinical characteristics, and temporal reasoning. To mitigate text-only shortcuts, we apply two-stage debiasing: (i) adversarial distractors are regenerated in a separate LLM call receiving only the stem and correct answer; (ii) a blind text-only stress-test identifies and reverts any newly introduced bias. Post-debiasing blind accuracy is 44.6% (prompted) and 37.1% (unprompted) vs. the 20% random baseline, with residual margins reflecting dataset priors rather than surface cues. Up to 10% of the VQA pairs were manually checked by the authors after debiasing with 98% correctness.

Baselines. Methods reported in this work include a set of Multimodal Large Language Models (MLLMs) such as different variants of GPTs, Gemini, Qwen,

Table 3. Colon-Bench Combined Results. Summary of model performance across the four benchmark tasks: visual prompted and unprompted video Visual Question Answering (VQA) accuracy (with visual box prompts and without them), binary lesion classification precision/recall/F1, and open-vocabulary video segmentation IoU/Dice. The video segemtnation is based on 3 box detections for each MLLM prompting the same EdgeTAM tracker [29]. Best scores per metric are shown in **bold**.

	VQA Accuracy		**Lesion Cls.**			**Segmentation**	
Model	Prompted	Unprompted	Precision	Recall	F1	IoU	Dice
CLIP [22]	-	-	34.5	**100**	51.3	-	-
ViCLIP [26]	-	-	34.8	98.5	51.4	-	-
Endo-CLIP [13]	-	-	41.9	95.2	58.2	-	-
Colon-ViCLIP [26]	-	-	49.0	84.2	62.0	-	-
SAM 3 [5]	-	-	-	-	-	2.5	2.9
GPT-4o	-	-	-	-	-	0.5	0.8
Claude Opus 4.6	-	-	-	-	-	16.1	20.5
GPT-5.2	-	-	-	-	-	30.7	36.5
GPT-5.4	-	-	-	-	-	34.5	41.1
Qwen3-VL 8B	32.9	38.3	34.4	**100**	51.2	10.4	13.1
Seed 1.6 Flash	38.1	45.4	**94.2**	24.3	38.6	2.6	3.5
Qwen-VL Max	39.1	45.4	0.0	0.0	0.0	25.6	29.6
Qwen3-VL 32B	39.3	44.4	0.0	0.0	0.0	12.7	15.9
Qwen3.5 Plus	44.3	60.5	36.2	24.6	29.3	16.7	21.0
Molmo 2-8B	46.1	53.4	52.9	46.7	49.6	2.6	3.8
Qwen3-VL 235B	46.7	56.6	22.2	0.7	1.4	13.6	16.9
Gemini 2.5 F. Li.	47.8	46.8	42.3	95.9	58.7	19.9	24.3
Qwen3.5 397B	49.0	60.1	10.0	0.4	0.7	16.6	21.0
GLM-4.6V	55.7	53.8	46.5	94.1	62.2	12.5	16.1
Seed 1.6	62.9	72.0	85.0	58.7	69.4	12.6	16.1
Gemini 3.1 F. Li.	69.2	67.7	72.6	90.8	**80.7**	37.4	43.4
Gemini 3 Flash	76.6	76.0	55.3	97.1	70.5	**48.3**	**54.7**
Gemini 3 Pro	**78.6**	**82.5**	66.1	93.0	77.3	45.0	51.3

Seed, GLM, and Molmo [21,2,8,10,12,11]. Also we report (for some corresponding benchmarks) SAM-3 [5], CLIP [22], Video CLIP (ViCLIP) [26] and its fine-tuned version on Colon-Bench text (Colon-CLIP), and a recent model Endo-CLIP [13]. Modles that do not have video API support (*E.g.* GPT-series) are only evluated on the segemntation task. All models in segemtnation (Except SAM-3) are based on 3-frame box detections prompting EdgeTAM tracker [29].

2.3 Colon-Bench Results

Table 3 reveals a clear performance hierarchy across all four Colon-Bench tasks. Gemini 3 Pro and Gemini 3 Flash consistently dominate, achieving the highest VQA accuracy on both prompted and unprompted splits and the best segmentation quality (IoU/Dice of 45-48%/51-55%), while Gemini 3.1 Flash Lite leads in classification F1 (80.7%). Among open-weight models, Seed 1.6 is the strongest overall, ranking third in VQA and delivering competitive classification F1 (69.4%)

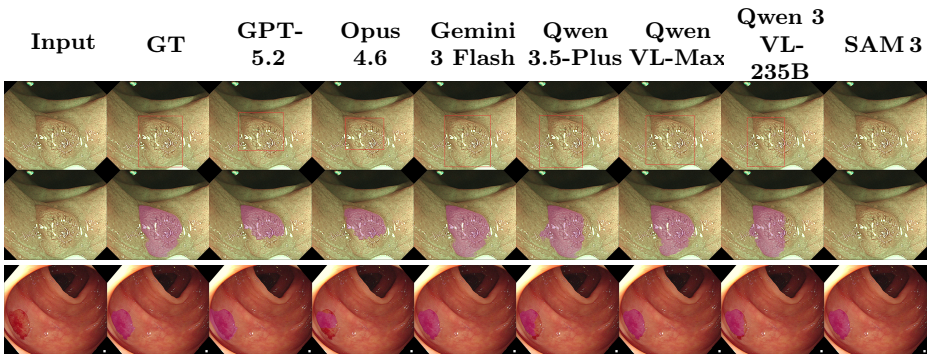

Fig. 3. Qualitative Comparisons. The first row shows detection (*Sessile Polyp*); the remaining rows show segmentation (*Sessile Polyp & Erythematous Region*). Columns show the input, ground truth, and model outputs. SAM-3 underperforms compared to modern MLLMs when paired with a prompted tracker such as EdgeTAM [29].

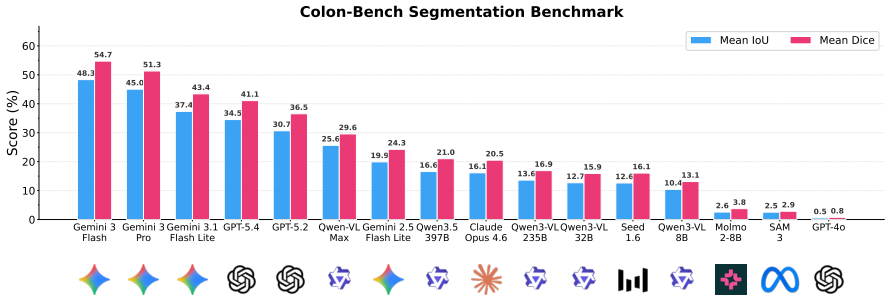

Fig. 4. Multi-Modal Large Language Models for Lesion Segmentation. Video object segmentation results on the 272-video benchmark, where MLLMs [21,2,8,10,12,11] first localize the target lesion with 3 bounding boxes and a EdgeTAM tracker [29] propagates masks.

with the highest precision among full-benchmark models (85.0%). The Qwen family shows mixed behavior: larger variants (Qwen3.5 397B, Qwen3-VL 235B) perform well on VQA yet nearly collapse on classification ($R < 1\%$, $F1 < 2\%$), suggesting they default to the majority (negative) benign class rather than detecting lesion presence. For segmentation, performance drops sharply beyond the top three models. Overall, our Colon-Bench demonstrates how these strong MLLM models can be used for lesion detection in short clips, beating specialized models like Endo-CLIP [13] by 30%. For the task of open-vocabulary lesion segmentation in video clips, an MLLM like GPT-5.4 (when combined with box-promptable EdgeTAM tracker [29]) beats SAM-3 [5] by 32.0% mIoU (see Fig. 4).

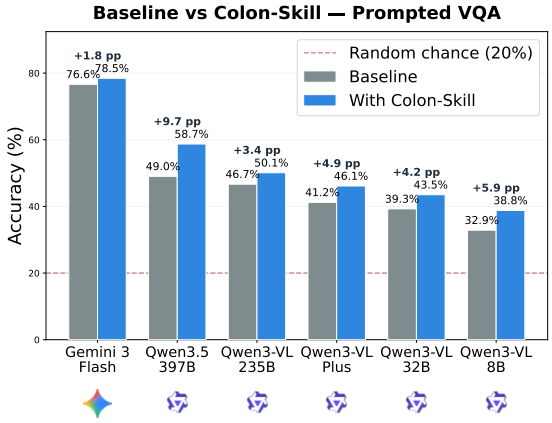

Fig. 5. Colon-Skill. Skill-augmented prompting on the prompted VQA split, where a distilled error-based skill context yields consistent accuracy improvements.

3 Analysis and Insights

3.1 Colon-Skill for MLLMs

To test whether Multimodal Large Language Models (MLLMs) benefit from structured domain knowledge at inference time, we use a two-stage skill-extraction and prompt-augmentation pipeline. First, we collect per-model VQA predictions on the full benchmark and stratify errors by lesion category (Fig. 2), retaining only questions that a majority of models answer incorrectly. These shared failure cases, along with their question/answer context and category metadata, are fed to a frontier large language model which synthesises a concise, natural-language *Colon-Skill*: morphological cues, common confusion traps, and a decision checklist tailored to colonoscopic VQA. Second, we prepend this Colon-Skill to every VQA prompt and re-evaluate all models under identical conditions, yielding consistent improvements up to +9.7% (Fig. 5). This suggests that distilling cross-model error patterns into structured textual guidance is an effective, training-free way to improve MLLM performance on specialised medical VQA tasks, particularly when models can integrate the added context.

3.2 Ablation Study

Effect of frame count on segmentation. Increasing the number of input frames for detection yields steady gains in downstream segmentation quality. For instance, expanding the context from 1 to 7 frames boosts the mean IoU of Gemini 3 Flash from 43.1% to 54.4% (mDice: 48.8% to 61.5%), and similarly improves Qwen-VL Max from 19.7% to 33.7% mIoU. However, this comes with a significant cost as these are per-window detections. We pick 3 equally-spaced frames per window as a good trade-off in all of our segmentation results.

Temporal context in VQA. Evaluating temporal context on the prompted VQA split showed that full video generally outperforms single-frame inputs, though the degree of impact is model-dependent. Restricting the input to a single frame reduced accuracy for Qwen3.5 397B (−7.0%), Qwen-VL Max (−3.4%), and Gemini 3 Flash (−2.5%). This highlights the importance of using videos instead of images used in previous polyp detection benchmarks.

4 Conclusions

We presented Colon-Bench, a novel agentic workflow and a comprehensive multi-task benchmark for full-procedure colonoscopy video understanding. By scalably and densely annotating 14 distinct lesion categories with masks, bounding boxes, and clinical text, we demonstrated that state-of-the-art MLLMs excel at high-level colonoscopy reasoning and spatial grounding with engineered context and agentic workflows.

Acknowledgments. We thank Dr. Jamal Hamdi, a surgeon with over 30 years of experience, who reviewed the accuracy of the Colon-Bench annotations together with the authors. This work was supported by the King Abdullah University of Science and Technology (KAUST) Center of Excellence for Smart Health. Additional support came from the KAUST Ibn Rushd Postdoctoral Fellowship.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ali, S., Jha, D., Ghatwary, N., Realdon, S., Cannizzaro, R., Salem, O.E., Lamarâo, D.V., Da Roza, C., Riegler, M.A., Halvorsen, P.: A multi-centre polyp detection and segmentation dataset for generalisability assessment. Scientific Data **10**(1), 75 (2023) 2
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 2, 6, 7
3. Biffi, C., Antonelli, G., Bernhofer, S., Hassan, C., Hirata, D., Iwatate, M., Maieron, A., Salvagnini, P., Cherubini, A.: Real-colon: A dataset for developing real-world ai applications in colonoscopy. Scientific Data **11**(1), 539 (2024). https://doi.org/10.1038/s41597-024-03359-0 2, 3, 4, 9
4. Bose, K., Kumar, A., Soundararajan, R., Mudgil, P., Ralmilay, S., Dutta, N., Singhal, M., Kumar, A., Sen, S., Patra, A., et al.: Multi-rads synthetic radiology report dataset and head-to-head benchmarking of 41 open-weight and proprietary language models. arXiv preprint arXiv:2601.03232 (2026) 4
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025) 6, 7

6. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. arXiv preprint arXiv:2405.19538 (2024) 2
7. Choudhuri, A., Gao, Z., Zheng, M., Planche, B., Chen, T., Wu, Z.: Polypsegtrack: Unified foundation model for colonoscopy video analysis. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Springer (2025) 2
8. Deitke, M., Clark, C., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025) 2, 6, 7
9. Fan, Y., Ma, X., Wu, R., Du, Y., Li, J., Gao, Z., Li, Q.: Videoagent: A memory-augmented multimodal agent for video understanding. In: European Conference on Computer Vision. pp. 75–92. Springer (2024) 2
10. Ge, Y., Zhao, Y., Zhu, Z., et al.: Planting a seed of vision in large language model. arXiv preprint arXiv:2307.08041 (2023) 2, 6, 7
11. Gemini Team: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024) 2, 6, 7
12. GLM Team: Glm-4: Open multilingual multimodal chat lms (2024), https://github.com/THUDM/GLM-4 2, 6, 7
13. He, Y., Zhu, Y., Fu, P., Yang, R., Chen, T., Wang, Z., Li, Q., Zhou, P., Yang, X., Wang, S.: Endo-clip: Progressive self-supervised pre-training on raw colonoscopy records. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Springer (2025), early accepted. Available at: https://arxiv.org/abs/2505.09435 2, 6, 7
14. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia Modeling. Lecture Notes in Computer Science, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37 2
15. Jong, M.R., Boers, T.G.W., Kusters, K.H.J., Jaspers, T.J.M., van der Putten, J.A., de With, P.H.N., van der Sommen, F., de Groof, A.J., Bergman, J.J.G.H.M., et al.: Gastronet-5m: A multicenter dataset for developing foundation models in gastrointestinal endoscopy. Gastroenterology **170**(1), 174–187 (2026). https://doi.org/10.1053/j.gastro.2025.07.030 2
16. Li, K., Fathan, M.I., Patel, K., Zhang, T., Zhong, C., Bansal, A., Rastogi, A., Wang, Z., Wang, G.: Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations. PLoS ONE **16**(8), e0255809 (2021) 2
17. Li, S., Ren, Y., Yu, Y., Li, H.: A survey of deep learning algorithms for colorectal polyp segmentation. Neurocomputing (2024) 2
18. Liu, S., Zheng, B., Chen, W., Peng, Z., Yin, Z., Shao, J., Hu, J., Yuan, Y.: Endobench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. Advances in Neural Information Processing Systems **38** (2026) 2
19. Ma, Y., Chen, X., Cheng, K., Li, Y., Sun, B.: Ldpolypvideo benchmark: A large-scale colonoscopy video dataset of diverse polyps. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference. pp. 387–396. Springer (2021) 2
20. Misawa, M., Kudo, S.e., Mori, Y., Maeda, Y., Kataoka, S., Ogata, N., Ichimasa, K., Kudo, T., Mori, K., Ishigaki, T.: Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointestinal Endoscopy **93**(4), 960–967 (2021) 2
21. OpenAI: Gpt-4 technical report (2023) 2, 6, 7

22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021) 2, 6
23. Song, Y., Zhang, Z., Wang, R., Zhong, L., Cai, C., Chen, J., Zhou, Y., Wang, X., Li, Z., Yang, L., Li, Z., Yan, H., Zhang, Q., Qian, D., Li, X.: Cas-colon: A comprehensive colonoscopy anatomical segmentation dataset for artificial intelligence development. Scientific Data **12**(1), 1382 (2025). https://doi.org/10.1038/s41597-025-05588-3 2
24. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **71**(3), 209–249 (2021). https://doi.org/10.3322/caac.21660 1
25. Tian, Q., Liao, H., Huang, X., Yang, B., Lei, D., Ourselin, S., Liu, H.: Endomamba: An efficient foundation model for endoscopic videos via hierarchical pre-training. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Springer (2025) 2
26. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., Luo, P., Liu, Z., Wang, Y., Wang, L., Qiao, Y.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. In: The Twelfth International Conference on Learning Representations (ICLR). OpenReview.net (2024), https://openreview.net/forum?id=MLBdiWu4Fw 2, 6
27. Wang, Z., Liu, C., Dou, Q., Zhang, S., et al.: Improving foundation model for endoscopy video analysis via representation learning on long sequences. IEEE Journal of Biomedical and Health Informatics pp. 3526–3536 (2025). https://doi.org/10.1109/JBHI.2025.3349925 2
28. Zeng, Q., Xie, Y., Lu, Z., Lu, M., Wu, Y., Xia, Y.: Segment together: A versatile paradigm for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging **44**(7), 2948–2959 (2025). https://doi.org/10.1109/TMI.2025.3360447 2
29. Zhou, C., Zhu, C., Xiong, Y., Suri, S., Xiao, F., Wu, L., Krishnamoorthi, R., Dai, B., Loy, C.C., Chandra, V., et al.: Edgetam: On-device track anything model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13832–13842 (2025) 3, 6, 7