Table of Contents with Clickable Links

Supplementary Table 1

Supplementary Table 1 Summary of key studies of NLPs in the field of gastrointestinal cancer.

Ref.	Year	Models	Objectives	Datasets	Results	Design
Harkema et al.[1]	2011	Rule-based NLP	Designed a rule-driven NLP system to evaluate colonoscopy quality standards.	679 colonoscopy and reports	89% accuracy and F1=0.724 for colonoscopy quality metrics	Retrospective
Denny et al.[2]	2012	Knowledge Map	Utilized the KnowledgeMap Concept Identifier to locate CRC tests in EHRs.	200 patients	Outperformed manual review with 93% recall for CRC tests vs. 71% for humans.	Retrospective
Mehrotra et al.[3]	2012	C-QUAL	Evaluated colonoscopy quality via the C-QUAL NLP framework.	24,157 colonoscopy reports	Kappa >0.67 quality measures, indicating substantial agreement.	Cross-sectional
Wagholikar et al.[4]	2012	NLP	Built an NLP-driven guidance system for colonoscopy surveillance intervals.	53 patients	Optimal colonoscopy surveillance recommendations in 96% of cases.	Retrospective
Imler et al.[5]	2013	CTAKES NLP	Applied the CTAKES for multidimensional classification of pathology findings.	500 oncology and pathology reports	93% accuracy for lesion sites and 80% for pathology landmarks	Retrospective
Imler et al.[6]	2014	CTAKES NLP	Built a CTAKES-driven model to determine optimal colonoscopy surveillance intervals.	10,798 colonoscopy reports	81% agreement with manual review, with Pearson R=0.833 for surveillance	Retrospective
Blumenth et al.[7]	2015	QPID NLP tool	Constructed a QPID-based predictive model for non-adherence to colonoscopy protocols.	1,531 patients	70.2% AUC for non-adherence prediction, with 92% specificity.	Retrospective
Raju et al.[8]	2015	Custom NLP	Developed custom NLP software for visualizing colonoscopy quality metrics.	12,184 colonoscopy patients reports	91.3% accuracy for screening, 99.9% identification accuracy	Retrospective

Sada et al.[9]	2016	ARC	Built an Automated Retrieval Console (ARC) to enhance liver tumor detection.	1,158 patients from LCKP-9	identified 87.97% of HCC cases, with 0.94 CPPV and 0.96 sensitivity	Retrospective
Nayor et al.[10]	2018	Custom NLP	Built a specialized NLP pipeline to calculate adenoma and serrated polyp detection rates.	8,032 screening colonoscopies	Perfect 100% precision and detection	Retrospective
Becker et al.[11]	2019	NLP pipeline	Developed a German NLP pipeline for guideline-compliant treatment evaluation	500 clinical notes from 23 patients	96.68% precision and 89.97% recall for tumor stage detection	Retrospective
Denny et al.[12]	2020	Knowledge Map	Applied KnowledgeMap to identify colonoscopy timing and procedural status.	323,944 patients	93% accuracy in CRC test, with 95% precision for diagnoses	Retrospective
Parthas et al.[13]	2020	NLP	Created an NLP-based diagnostic model for serrated polyposis syndrome.	200 patients	Diagnosis recall: 67.03%, precision: 95.09%	Retrospective
Fevrier et al.[14]	2020	SAS YABE NLP	Used the SAS YABE! tool for standardized data extraction from colonoscopy reports.	401,566 reports with pathology reports	Cohen's κ of 93.9–99% and >97% PPV for common diagnostic classifications	Retrospective
Li et al.[15]	2021	ML and NLP fusion	Fused machine learning and NLP to predict Lynch syndrome in MSI-H patients.	5,520 patients	perfect 100% sensitivity, specificity, PPV, and NPV for disease detection	Retrospective
Bae et al.[16]	2022	smartTA	Integrated regular expressions with smartTA templates for indicator assessment.	2,626 colonoscopy reports	NLP pipeline reached 90~100% accuracy for polyp subtypes and lesions	Retrospective
Song et al.[17]	2022	Custom NLP	Developed a custom NLP pipeline to parse upper gastrointestinal endoscopy reports.	1,000 validation, 248,866 applications	96.6% sensitivity, PPV, accuracy, and F1 for EGD report extractions	Retrospective
Li et al.[18]	2022	ENDANGEL - AS NLP	Created the ENDANGEL-AS system for identification and surveillance planning.	22,208 patients across various texts	100% internal accuracy and 99.93% external accuracy in high-risk patient	Retrospective
Laigle et al.[19]	2023	OCR and NLP	Combined OCR and NLP for structured extraction of healthcare quality metrics.	35,314 colonoscopy reports	95% accuracy for most clinical variables, with some metrics exceeding 99%.	Retrospective

Seong et al.[20]	2023	Bi-LSTM-CRF	Implemented a Bi-LSTM-CRF architecture for medical report information extraction.	280,668 colonoscopy reports	F1 scores >0.95 across diverse report findings.	Retrospective
Tinmouth et al.[21]	2023	NLP	Used NLP for automated adenoma detection rate (ADR) calculation.	1,461 pathology reports	99.6% sensitivity and 99.81% specificity for adenoma detection.	Retrospective
Wenker et al.[22]	2023	cLamp	Applied the cLamp system to identify dysplasia in Barrett's Esophagus.	1,000 patients for NLP validation	98.7% accuracy, 100% precision, and 92.2% recall for Barrett's dysplasia.	Retrospective
Ganguly et al.[23]	2023	NLP	Developed an NLP solution for adenoma detection and automated report generation.	2,276 colonoscopy procedures	100% sensitivity, specificity, and accuracy for detection and report generation.	Retrospective
Benson et al.[24]	2023	NLP	Created an NLP system to support colorectal polyp surveillance decisions	26,434 pathology reports	90.8% precision, 94.9% recall, and an F1-score of 98.4%	Retrospective
Li et al.[25]	2023	NLP	Integrated NLP and machine learning for liver metastasis risk prediction	3,641 colorectal Liver patients	high accuracy in liver metastasis risk prediction	Retrospective

Abbreviations: CRC: Colorectal cancer; HCC: Hepatocellular carcinoma; IBS: Irritable bowel syndrome; IBD: Inflammatory bowel disease; LSTM: Long Short-Term Memory; NLP: Natural language processing; MoE: Mixture of experts; LLMs: Large language models; SNP: Single nucleotide polymorphism; TCGA: The Cancer Genome Atlas.

Supplementary Table 2

Supplementary Table 2 Methodological details of Vision Foundation Models-assisted endoscopy.

Model	Country	Dataset sizes	Evaluation metrics	Fine-tuning strategies	Performance benchmarks	GPU
Surgical-DINO[76] [a]	China	SCARED dataset contains 35 endoscopic videos with 22950 frames. Split into 15351 for training, 1705 for validation, and 551 for testing. Hamlyn dataset has 21 videos for validation.	Abs Rel, Sq Rel, RMSE, RMSE log, δ	LoRA layers added to DINOv2. Freezing the DINO image encoder and only optimizing the LoRA layers and depth decoder.	Surgical-DINO achieved Abs Rel: 0.053, Sq Rel: 0.377, RMSE: 4.296, RMSE log: 0.074, δ: 0.975 on SCARED dataset.	1 RTX 3090
ProMISe [77]	China	Kvasir: 1450 images; EndoScene: 785 images; ColonDB: 112 images; ETIS: 144 images; ISIC2018: 900 images	mDice, Dice Score, IoU	No fine-tuning. Only APM and IPS modules are trained while keeping SAM parameters frozen.	Average mDice: 0.915 (ProMISe on ISIC2018), 0.874 (ProMISe on ColonDB), 0.854 (ProMISe on ETIS)	1 A100
Polyp-SAM[78]	USA	Kvasir: 1,000 images; CVC-ClinicDB: 612 images; CVC-ColonDB: 380 images; ETIS: 196 images; CVC-300: 60 images	Dice Similarity Coefficient (DSC), mean Intersection over Union (mIoU)	Two strategies: Pretrain only the mask decoder while freezing all encoders; Pretrain the image encoder, prompt encoder, and mask decoder	CVC-ColonDB (89.4% DSC), CVC-300 (92.4% DSC), and ETIS (90.3% DSC). Polyp-SAM-L achieved state-of-the-art results in CVC-300 (92.9% DSC) and ETIS (90.5% DSC)	2 V100
Endo-FM[79]	China	Pre-train: 32896 clips, 5.02M frames; Downstream: 335 clips, 506005 frames	F1 score (%) for PolypDiag, Dice (%) for CVC-12k, F1 score (%) for	PolypDiag: linear layer added; CVC-12k: TransUNet with Endo-FM backbone; KUMC: STFT with Endo-FM	Endo-FM: 90.7±0.4 (F1), 73.9±1.2 (Dice), 84.1±1.3 (F1); Scratch: 83.5±1.3, 53.2±3.2, 73.5±4.3; VCL: 87.6±0.6, 69.1±1.2, 78.1±1.9	NA

			KUMC	backbone		
ColonGPT[80]	China	303,001 colonoscopy images for training, 450,724 human-machine dialogues for tuning	Accuracy for classification tasks (CLS), IoU for REC	Two-stage training: pre-alignment with image-caption pairs, followed by supervised fine-tuning with CLS, REG, and REC data using LoRA	Achieved highest scores in CLS (94.06%), REG (99.96%), and REC (85.74%) tasks on validation set; superior generalisation ability on unseen samples	2 H200
DeepCPD[81]	India	PolypsSet: 35,981 images (25,187 polyp, 10,794 non-polyp); CP-CHILD-A: 938 images (549 polyp, 389 non-polyp); CP-CHILD-B: 1,000 images (500 polyp, 500 non-polyp); Kvasir V2: 2,000 images (1,000 polyp, 1,000 non-polyp).	Recall, Precision, F1-score, Accuracy, Specificity, MCC	The model is initialized with pre-trained ViT weights. Hyperparameters are optimized for colonoscopy datasets, including Adam optimizer	DeepCPD achieved accuracy >98.05%, recall >98.10%, precision >97.71%, F1-score >97.80%, and specificity >97.00% across all datasets	2 V100
OneSLAM[82]	USA	Sinus (n=15), Colon (n=20), Joint (n=12), Laparoscopy (n=18)	ATE (Absolute Trajectory Error), RPE (Relative Pose Error), RMSE	Zero-shot adaptation using TAP model	ATE: 1.2-3.5 mm; RPE: 1.8-4.0 deg; P2M RMSE: 0.8-2.1 mm; P2P RMSE: 0.6-1.9 mm	NA
EIVS[83]	China	Training: 613 WLE images, 637 ; Validation: 204 WLE images, 211 chromoendoscopy images; Test set: 204 WLE images, 211 chromoendoscopy images	CLIP embeddings and Maximum Mean Discrepancy distance; Gaussian RBF kernel	Unsupervised Cycle-Consistency	1EIVS: Parameters=63.41M, Time=0.0151s per image, CMMD×100=15.99.	NA
APT[84]	China	Kvasir-SEG: 1000 images (888	Dice, mIoU, Recall,	Parameter-efficient fine-	Dice scores of 91.64% (Kvasir-SEG),	2

		train, 112 test); CVC-ClinicDB: 612 images (535 train, 77 test); EndoTect: 200 images (175 train, 25 test)	Precision	tuning: only adapters and prompt generator are trained; SAM backbone frozen; Adam optimizer	95.08% (CVC-ClinicDB), and 92.57% (EndoTect); mIOU of 86.93%, 90.86%, 88.72%; Precision of 93.06%, 96.08%, 95.18%	RTX 3090
FCSAM[85]	China	Gastric cancer dataset: 630 pairs, Kvasir-SEG: 1000 images, CVC-ClinicDB: 612 images	IoU, Dice coefficient, Accuracy	LNLoRA fine-tuning strategy (LayerNorm and low-rank-based)	IOU: 79.3%, Dice coefficient: 88.1%, Accuracy: 93.2%	1 RTX 4090
Dua-PS Net[86]	China	Kvasir-SEG: 1000 images (800 train, 100 valid, 100 test); CVC-ClinicDB: 612 images (490 train, 61 valid, 61 test); CVC-ColonDB: 380 images; ETIS-LaribPolypDB: 196 images; CVC-300: 60 images	mDice, mIoU, Fwβ, Sα, mEφ, MAE	Fine-tuning conducted using transfer learning with pre-trained PVTv2-B3 on ImageNet. Model trained on CVC-ClinicDB/Kvasir-SEG	CVC-ClinicDB: mDice 0.9514, mIoU 0.9083; Kvasir-SEG: mDice 0.9253, mIoU 0.8746; Cross-dataset testing showed strong generalization, ETIS: mDice 0.868, mIoU 0.792	1 RTX 3060
EndoDINO[87]	USA	Pre-trained on 3.5B frames from 130k+ endoscopy videos; curated to 100K–10M images. Evaluated on HyperKvasir (~10.6K) and LIMUC (11,276).	Macro/Micro F1, mDice/mIoU/mPrec/mRec, Macro F1/AUROC.	DINOv2 methodology, Downstream adaptation uses linear probing, training only a simple linear classifier while keeping the backbone frozen.	SOTA across tasks: Macro F1 0.995 (1% data), mIoU 0.864 (polyp seg), AUROC 0.942, Macro F1 0.715 (4-class Mayo).	8 H100
PolypSegTrack[88]	USA	Kvasir-SEG (900 images), CVC-ClinicDB (550 images), PolypDB, PolypGen, KUMC (28k images), ETIS, CVC-ColonDB, CVC-300	Dice score, IoU score, Precision, Recall, F1 score	One-step fine-tuning on colonoscopic videos without first pre-training on colonoscopic videos	State-of-the-art performance on multiple polyp benchmarks, outperforming existing methods in detection, segmentation, classification	NA
AiLES[89]	China	100 patients, 5111 frames (4130 for development, 981 for	Dice, IoU, Recall, Specificity,	AiLES not fine-tuned from external model	Dice: 0.76±0.17, IOU: 0.61±0.19, Recall: 0.73±0.21, Specificity:	2 RTX

		independent test set)	Accuracy, Precision		0.99±0.01, Accuracy: 0.99±0.01, Precision: 0.79±0.16	3090
PP-SAM [90]	USA	Kvasir (1000 images), ClinicDB (612 images), EndoScene (60 images), ColonDB (379 images)	DICE similarity score	Fine-tuning with variable bounding box prompt perturbations	1-shot fine-tuning on Kvasir with 50-pixel BBP perturbations during inference boosts DICE score by 20%	RTX A6000
SPHINX-Co[91]	China	CoPESD includes 17,679 images	GPT-4 score, mIoU, Accuracy and F-score	Fine-tuned SPHINX-X model on CoPESD dataset with cosine learning rate scheduler	Achieves GPT score of 83.98 and mIoU of 70.48 with full dataset.	NA
LLaVA-Co[91]	China	CoPESD includes 17,679 images	GPT-4 score, mIoU, Accuracy and F-score	Fine-tuned LLaVA-1.5 model on CoPESD dataset with cosine learning rate scheduler.	Achieves GPT score of 85.63 and mIoU of 60.23 with full dataset.	NA
ColonCLIP[92]	China	Total: 338,671 images; Base Dataset: 300,841 training + 35,288 testing; Incremental Dataset: total 2,542	Recall, F1-score, Precision	Freeze CLIP encoders, train text and visual prompts; Freeze prompts, finetune CLIP encoders	Base Phase: F1=10.65%; Adaptive Phase: Recall=12.32%, F1=6.02%, Precision=6.05%	NA
PSDM[93]	China	PolypGen: 1,537 images; ETIS: 80; CVC-ClinicDB: 612; CVC-ColonDB: 300; CVC-300: 140; Kvasir: 2,000; Polyplus: 100; Malignant Polyp Dataset: 462.	mDice, mIoU; F1 score, mAP50, mAP50-95	Continual learning with prompt replay to incrementally train on multiple datasets. Trained from scratch with compositional prompts.	PolypGen: mDice = 76.70 (vs. 74.27 baseline), mIoU = 69.89 (vs. 67.92); F1 score +2.12%, mAP50 +3.09%. detection: mAP50-95 increased from 57.74 to 60.83 with YOLOv5.	NA
PathoPolyp-Diff[94]	India/ USA	49,136 polyp frames (SUN Database); 109,554 non-polyp frames; ISIT-UMR Colonoscopy Dataset with adenomatous and	KID, Precision, Recall, F1-score, Balanced Accuracy	Fine-tuned Stable Diffusion v1-4 with text prompts for polyp/non-polyp and quality; Locked first U-Net block, fine-	Balanced Accuracy improvement of 7.91% in classification; Cross-class label learning improved video-level analysis by 18.33%; Lowest KID:	A100 /Titan-Xp

	hyperplastic polyps	tuned remaining blocks with prompts.	0.036 at 1,000 iterations; Best F1-score at 10,000 iterations with KID 0.0605

[a]: Reference numbers here indicated reference in the main text.

Abbreviations: ViT: Vision Transformer; Abs Rel: Absolute Relative Error; Sq Rel: Squared Relative Error; RMSE: Root Mean Square Error; RMSE log: Logarithmic Root Mean Square Error; δ: the accuracy with a given threshold; LoRA: Low-Rank Adaptation; DINOv2: Data-efficient Image Transformer version 2; mDice: mean Dice score; IoU: Intersection over Union; APM: a module in the discussed method; IPS: another module in the discussed method; SAM: Segment Anything Model; DSC: Dice Similarity Coefficient; F1 score: a measure of test accuracy; CLS: Classification task; REG: Regression task; REC: Referring Expression Comprehension; MCC: Matthews Correlation Coefficient; ATE: Absolute Trajectory Error; RPE: Relative Pose Error; P2M RMSE: Point-to-Mesh Root Mean Square Error; P2P RMSE: Point-to-Point Root Mean Square Error; CMMD: Cross-Modal Maximum Mean Discrepancy; LNLoRA: LayerNorm and low-rank-based fine-tuning strategy; MAE: Mean Absolute Error; GPT: Generative Pre-trained Transformer; mAP50 and mAP50-95: mean Average Precision at 50% and from 50% to 95% Intersection over Union thresholds respectively; KID: Kernel Inception Distance.

Supplementary Table 3

Supplementary Table 3 Vision Foundation Models benchmarked in endoscopy.

Model	Year	Country	Base Model	Training Algorithm	Parameters	Datasets	Benchmarked in Ref.[a]	Model Type	GPUs	Source Code link
TimeSformer[26]	2021	USA	ViT	ImageNet Pretraining	121.4M	Kinetics-400, Diving-48 etc.	Endo-FM[79]	Vision	416 V100 hours	https://github.com/facebookresearch/TimeSformer
ST-Adapter[27]	2022	China	ViT-B	Fine-tuning	82M	Epic-Kitchens-100, Something	Endo-FM[79]	Vision	184 V100 hours	https://github.com/linziyi96/st-adapter
PolypsAlign[28]	2021	UK	ViT-S	Vision Transformer	NA	Kvasir, Nerthus	ColonGPT[80]	Vision	NA	https://github.com/qinwang-ai/PolypsAlign
CoTracker[29]	2024	USA	Transformer	Unrolled Training	NA	DAVIS, TAP-Vid-Kubric etc.	OneSLAM[82]	Vision	32 A100	https://github.com/facebookresearch/CoTracker
ECTransNet[30]	2023	China	Transformer	Transformer	NA	multiple	Dua-PSNet[86]	Vision	RTX3090	NA
PolypPVT[31]	2023	China	Transformer	Transformer	NA	multiple	Dua-PSNet[86]	Vision	NA	https://github.com/DengPingFan/PraNet
MCSFNet[32]	2024	China	Transformer	Transformer	NA	multiple	Dua-PSNet[86]	Vision	2 RTX3090	https://github.com/WYJGR/MCSF-Net
Etrolizumab[33]	2023	USA	ViT-B	DINOv1 (SSL)	NA	Etro Dataset	EndoDINO[87]	Vision	8 A100	NA
ArgesF	2024	USA	ViT-B	DINOv2 (SSL)	86M	UNIFI,	EndoDINO[	Vision	4 A10G	NA

M[34]						JAKUC etc.	87]			
DINOv2 [35]	2024	USA	ViT-g/14	DINOv2 (SSL)	1B	LVD-142M	PP-SAM[90]	Vision	20 A100	https://github.com/facebookresearch/dinov2
YOLO World	2024	China	YOLOv 8, RepVL	Vision-Language Pre-training	48M	Objects365V1, GQA, Flickr30k etc.	PP-SAM[90]	Vision	1 V100	https://github.com/AILab-CVC/YOLO-World
GroundingDINO [36]	2024	USA	ViT (DINO)	Grounded Pre-Training (SSL)	172M	COCO, O365, LVIS, etc.	PP-SAM[90]	Multimodal	16 V100	https://github.com/IDEA-Research/GroundingDINO
SAM[37]	2023	USA	ViT-H	MAE Pretraining	3B	SA-1B (11M images)	PP-SAM[90]	Vision	256 A100	https://segment-anything.com/
MedSAM[38]	2024	Canada	ViT-B	Fine-tuning	93.7M	Public data	PP-SAM[90]	Vision	20 A100	https://github.com/bowang-lab/MedSAM
GPT-4[39]	2023	USA	Transformer	RL with Human Feedback	1.8T	Web data, Books, etc.	ColonCLIP[92]	Multimodal	10,000+ A100	No
Claude-3-Opus	2024	USA	Transformer	RL with Human Feedback	NA	NA	ColonCLIP[92]	Multimodal	NA	NA
Gemini-1.5[40]	2024	USA	Transformer	RL with Human Feedback	NA	NA	ColonCLIP[92]	Multimodal	NA	NA
CLIP[41]	2021	USA	ViT-L4	Contrastive Learning	3.2B	WebImageText (WIT)	ColonCLIP[92]	Multimodal	592 V100	https://github.com/openai/CLIP
BiomedCLIP[42]	2025	USA	ViT-B	CLIP	86M	PMC-15M	ColonCLIP[92]	Multimodal	16 A100, V100	https://github.com/microsoft/BiomedCLIP

Stable diffusion[43]	2022, v1-4	USA	Transformer	DDPM	1.4B	LAION-400M, WebDoc etc.	PathoPolyp-Diff[94]	Generative	8 A100	https://github.com/Stability-AI/StableDiffusion

[a]: Reference numbers here indicated reference in the main text.

Abbreviations: ViT: Vision Transformer; SSL: Self-Supervised Learning; MAE: Masked Autoencoder; RL: Reinforcement Learning; DDPM: Denoising Diffusion Probabilistic Models; GQA: Google Questions Answering; LVIS: Large Vocabulary Instance Segmentation; SA-1B: Segment Anything 1-Billion mask dataset; PMC: PubMed Central; WIT: WebImageText; LAION: Large-scale Artificial Intelligence Open Network.

Supplementary Table 4

Supplementary Table 4 Methodological details of Vision Foundation Models-assisted radiology.

Model	Country	Dataset sizes	Evaluation metrics	Fine-tuning strategies	Performance benchmarks	GPU
PubMedCLIP[98] [a]	Germany	ROCO: >80K; VQA-RAD: 315 images, 3,515 QA pairs; SLAKE: 642 images, >7,000 QA pairs	Overall accuracy, open-end accuracy, closed-end accuracy	Fine-tuned on ROCO dataset, Adam optimizer	PubMedCLIP achieves up to 3% improvement in overall accuracy	NA
RadFM[97]	China	MedMD: 16M image-text pairs, including 15.5M 2D images and 500k 3D scans with captions or diagnosis labels. RadMD: 3M multi-modal samples	For classification tasks: Accuracy, F1 score. For open-ended tasks: BLEU, ROUGE, BERT-sim.	Pre-trained on MedMD and subsequently fine-tuned on RadMD.	RadFM outperforms existing multi-modal foundation models	32 A100
Merlin[99]	USA	Clinical dataset: 6,387,231 2D images from 15,331 CTs for training; 2,099,217 images from 5,060 CTs for validation; 2,142,061 images from 5,137 CTs for testing	Dice score, F1 score, AUROC, AUPRC. Radiology report generation: BLEU score, ROUGE-2, BERT score, RadGraph-F1	Multi-task learning with EHR and radiology reports. Fine-tuning for specific tasks such as 5-year disease prediction and 3D semantic segmentation.	Internal F1 score 0.741, External F1 score 0.647. Macro-average AUROC 0.812. Multi-disease 5-year prediction: AUROC 0.757.	A6000
Med-Gemini[100]	USA	MedQA (1273 questions), NEJM CPC (303 cases), GeneTuring (600 QA pairs), Clinical Abstraction (81 samples)	Accuracy for MedQA, Top-1 and Top-10 accuracy for NEJM CPC, Averaged accuracy for GeneTuring	Instruction fine-tuning Gemini 1.0/1.5 on medical QA, multimodal and long-context corpora	91.1% accuracy on MedQA, 72.3% accuracy on NEJM CPC, 86.0% accuracy on GeneTuring	NA
HAI-DE	USA	Over 500,000 cases across different	ROC AUC and accuracy for	Fine-tuning on downstream	Achieves high performance	NA

F[101]		anatomic regions	classification tasks	tasks with limited labeled data	across diverse tasks	
CT-FM[102]	USA	Imaging Data Commons: 148,000 CT scans, TotalSegmentator: 1,228 CT scans annotated with 117 structures, SinoCT: 9,779 head CT scans, CQ500: 491 head CT scans for zero-shot evaluation, OrganMNIST3D: 1,743 abdominal organ samples for retrieval tasks	Dice score for segmentation tasks, F1 score and AUC-ROC for classification tasks, and average precision, hit rate, and F1 score for retrieval tasks.	Trained from scratch using a self-supervised learning strategy on a large-scale dataset of 148,000 CT scans from the Imaging Data Commons, employing a SegResNet encoder for the pre-training phase.	mean Dice coefficients of 0.9058 for whole-body segmentation, outperform in tumor segmentation tasks, high F1 scores and AUC-ROC values in head CT triage classification.	4 RTX 8000
MedVersa[103]	USA	MedInterp, comprises 91 publicly available datasets encompassing 29 million instances across 11 different tasks and seven imaging modalities.	BLEU-4, BertScore, CheXbert, RadGraph, RadCliQ for report generation; F1, IoU, DICE for classification, detection, segmentation evaluation.	Trained from scratch on MedInterp with visual-linguistic supervision; adaptable to diverse medical imaging tasks.	Achieves SOTA in nine tasks, >10% better than specialized models; matches or exceeds human reports in 71% of cases.	24 A100
iMD4GC[104]	China	GastricRes: 698 patients (240 with complete data); GastricSur: 801 patients (456 with complete data); TCGA-STAD: 400 patients.	Treatment response: AUC (primary), accuracy, precision, recall, F1-score. Survival analysis: C-index, time-dependent AUC.	Pre-trained encoders (ResNet-50 for CT, CTransPath for WSI); fusion architecture with cross-modal interaction and knowledge distillation; trained from scratch.	GastricRes: AUC 80.2%, acc 74.8%, prec 75.1%, rec 74.5%, F1 74.8%. GastricSur: C-index 71.4%. TCGA-STAD: C-index 66.1%.	8 RTX 3090
Yasaka et al[105]	Japan	5194 training images (927 breast carcinoma, 2180 esophageal carcinoma, 2087 no lesion); 583 validation (80 breast, 233	AUC for diagnostic performance; sensitivity and accuracy for breast and esophageal carcinoma	Fine-tuned on CT images with text labels ("suspicious of breast carcinoma"/"esophageal carcinoma"/"no lesion"); used	AUC for breast carcinoma: 0.890 (95%CI 0.871–0.909), AUC for esophageal carcinoma:	Quadro P5000

	esophageal, 270 no lesion); 7349 testing (184 breast, 246 esophageal, 6919 no lesion).	detection.	LoRA with fc1 layer tuning in vision and q-former; single GPU, desktop setup.	0.880 (95%CI 0.865–0.894)

[a]: Reference numbers here indicated reference in the main text.

Abbreviations: QA: Question Answering; ACC: Accuracy; F1: F1 score; AUROC: Area Under the Receiver Operating Characteristic Curve; AUPRC: Area Under the Precision-Recall Curve; BLEU: Bilingual Evaluation Understudy; ROUGE: Recall-Oriented Understudy for Gisting Evaluation; BERT: Bidirectional Encoder Representations from Transformers; RadGraph: Radiology Graph-based metric; Dice: Dice Similarity Coefficient; IoU: Intersection over Union; DSC: Dice Similarity Coefficient; ASD: Average Surface Distance; AUC: Area Under the Curve; C-index: Concordance index; WSI: Whole Slide Image; LoRA: Low-Rank Adaptation.

Supplementary Table 5

Supplementary Table 5 Vision Foundation Models benchmarked in radiology.

Model	Year	Country	Architecture	Training Algorithm	Parameters	Datasets	Benchmarked in Ref.[a]	Model Type	GPU	Source Code link
OpenFlamingo[44]	2023	USA	ViT-L/14	CLIP	3B~9B	LAION-2B, C4 etc.	RadFM[97]	Multimodal	64 A100	https://github.com/mlfoundations/open_flamingo
MedFlamingo[45]	2023	USA	OpenFlamingo-9B	In-context Learning	8.3B	Multiple	RadFM[97]	Multimodal	8 A100	https://github.com/snap-stanford/med-flamingo
GPT-4V[46]	2023	USA	Transformer	Multimodal Fusion	1.8T	NA	RadFM[97]	Multimodal	NA	NA
OpenCLIP[47]	2023	USA	ViT-H/14	CLIP	600M	LAION-2B	Merlin[99]	Multimodal	1520 A100	https://github.com/mlfoundations/open_clip
BiomedCLIP[42]	2023	USA	ViT-B/16, PubMedBERT	OpenCLIP	NA	PMC-15M	Merlin[99]	Multimodal	16 A100, V100	https://aka.ms/biomedclip
MulT[48]	2019	USA	Transformer	Crossmodal Attention	200K	CMU-MOSI, MOSEI etc.	iMD4GC[104]	Multimodal	GTX1080 Ti	https://github.com/yaohungt/Multimodal-Transformer
Performer[49]	2021	USA	Transformer	PG-19	NA	TrEMBL	iMD4GC[104]	Multimodal	1 V100	NA
Nystromformer[50]	2021	USA	Transformer	Nystrom-based	NA	Wikipedia, BookCorpus	iMD4GC[104]	Vision	8 V100	https://github.com/mlpen/Nystromformer
COM[51]	2023	China	Transformer	Contrastive	12.28	CUB, ORL,	iMD4GC[104]	Multim	1	https://github.com/your-

	Learning	M	PIE, YaleB		odal	RTX3090	repo-link

[a]: Reference numbers here indicated reference in the main text.

Abbreviations: ViT: Vision Transformer; CLIP: Contrastive Language-Image Pre-training; CoCa: Contrastive Captioners; LAION: Large-scale Artificial Intelligence Open Network; CUB: Caltech-UCSD Birds-200-2011 dataset; ORL: Olivetti Research Laboratory face database; PIE: Pose, Illumination, and Expression database; YaleB: Extended Yale Face Database B; CMU-MOSI: Carnegie Mellon University Multimodal Opinion Sentiment Intensity dataset; MOSEI: Multimodal Opinion Sentiment and Emotion Intensity dataset; LoRA: Low-Rank Adaptation.

Supplementary Table 6

Supplementary Table 6 Methodological details of Vision Foundation Models-assisted pathology.

Model	Dataset sizes	Evaluation metrics	Fine-tuning strategies	Performance benchmarks	GPU
LUNIT-SSL[110]a	Pre-trained on 32.6M pathology image patches (19M from TCGA, 13.6M from TULIP) extracted from over 36,000 WSIs at 20x and 40x magnifications.	Evaluated using Top-1 accuracy for classification; mPQ for nuclei instance segmentation.	Trained via self-supervised learning (MoCo v2, SwAV, Barlow Twins, DINO) on unlabeled pathology patches; evaluated by linear probe (frozen backbone) or full fine-tuning on downstream tasks.	Sets new benchmarks; outperforms ImageNet-pretrained models on pathology tasks (BACH, CRC, PCam, MHIST classification; CoNSeP segmentation), especially in low-label settings.	64 V100
CTransPath[111]	Evaluated on TCGA, NCT-CRC-HE, and other datasets; tested with 0.5%, 1%, 10%, 50%, 100% labeled data to assess performance in low-label scenarios.	Evaluated using ACC@1, ACC@3, ACC@5, mMV@5, F1, AUC; t-SNE for feature discrimination visualization.	Trained with self-supervised SRCL loss on histopathology patches; for downstream tasks: frozen backbone with linear/SVM classifier or feature extraction only.	Achieves SOTA, surpassing SSL (e.g., SimCLR, MoCo v3) and ImageNet-pretrained models; +0.6% ACC over KimiaNet ensemble on CRC; outperforms fully supervised baseline with only 1% labeled data on NCT-CRC-HE.	48 V100
Phikon[112]	The models were pre-trained on histology tile datasets ranging from 4 million tiles specific to colorectal cancer (TCGA-	Model performance was evaluated using slide-level ROC AUC for classification tasks, Harrell C-index for overall	The models were trained using the iBOT self-supervised learning framework with masked image modeling on histology tiles and subsequently	The iBOT PanCancer model achieved state-of-the-art performance, outperforming models pre-trained on ImageNet, with MoCoV2, or with iBOT on	NA

	COAD) to a large-scale pan-cancer dataset comprising 40 million tiles.	survival prediction, and patch-level accuracy, F1, and ROC AUC scores.	fine-tuned for downstream tasks using multiple instance learning (MIL) algorithms.	smaller/colorectal-specific datasets across various weakly-supervised histopathology tasks.	
REMEDIS[113]	The model leverages unlabeled datasets ranging from ~200K (dermatology) to ~2.3M (fundus images), alongside labeled ID/OOD splits in the tens of thousands.	Each task is assessed using AUC for binary tasks and top-3 accuracy for multi-class tasks, computed on both in-distribution (ID) and out-of-distribution (OOD) test sets.	Supervised-pretrained on JFT-300M natural images, then contrastive-self-supervised on unlabeled medical images, and finally end-to-end fine-tuned on labeled ID data plus optional fractions of labeled OOD data.	REMEDIS surpasses strong ImageNet/JFT baselines by up to 11.5 % ID gain and 10.7 % OOD gain, achieving expert-level performance with only 1–33 % of retraining data.	256 Google Cloud TPU
Virchow[114]	Trained on ~1.5M H&E WSIs from ~100K MSKCC patients; 4–10× larger than prior pathology model datasets.	Evaluated primarily by AUC; also specificity, sensitivity; significance via DeLong's, Cochran's Q + McNemar's with correction.	632M-parameter ViT trained with DINOv2 on WSI tiles; downstream: tile embeddings fed into lightweight aggregator networks for specimen-level tasks.	Achieves SOTA: highest or statistically tied AUC across most common and rare cancers; superior generalization to external data, strong in rare cancer and biomarker prediction.	NA
Virchow2[115]	Trained on 3.1M WSIs from diverse institutions (e.g., MSKCC), covering multiple tissue types and staining protocols.	Performance was evaluated using weighted F1 score, AUROC, and AP across 8 to 12 tile-level benchmark tasks on public datasets.	It was trained self-supervised with DINOv2 on histopathology tiles, and fine-tuned using linear probes or full-tuning on downstream classification tasks.	Achieved highly performance across multiple computational pathology benchmarks, especially in mitosis detection and rare cancer classification.	512 V100
Virchow2G[115]	Trained on same 3.1M WSI dataset as Virchow2; diverse tissue types and	Evaluated using weighted F1, AUROC, AP on tile-level classification and	Trained self-supervised with DINOv2 on pathological images and fine-tuned using linear	Virchow2G achieved state-of-the-art performance on several benchmarks, particularly excelling in tasks	512 V100

	staining methods.	detection tasks.	probes or full fine-tuning on downstream tasks.	requiring fine-grained detail like mitosis detection.	
Virchow2G mini[115]a	Trained on distilled dataset from same 3.1M WSIs as Virchow2G; leverages knowledge from larger teacher model.	Evaluated with weighted F1, AUROC, AP on tile-level classification benchmarks.	Obtained by knowledge distillation from the Virchow2G teacher model, then fine-tuned on downstream tasks using linear probes or full fine-tuning.	Virchow2G Mini outperformed larger models like H-optimus-0 and GigaPath on several benchmarks while being significantly smaller in parameter count.	256 V100
UNI[9]	Evaluated on 345,021 ROIs (PRAD), 55,360 ROIs (pan-cancer), 2,399 ROIs (colorectal polyp), and up to 1,620 whole slides (OT-43, OT-108).	Assessed by macro AUROC and Top-1 accuracy; quadratic weighted κ for ISUP grading.	Trained with DINOv2 self-supervised learning on diverse histology images; used for downstream tasks via few-shot learningor linear evaluation, not full fine-tuning.	UNI shows superior label efficiency: median 8-shot performance exceeds 128/256-shot of others; lowest few-shot sometimes surpasses best ResNet-50, CTransPath, REMEDIS.	32 A100
Phikon-v2[116]	The model was pre-trained on 58,359 WSIs (456 million tiles) from >100 public cohorts covering >30 cancer sites.	Slide-level AUC (area under ROC) on external validation cohorts, reported with 95% CIs and permutation tests.	Pre-trained with DINOv2, then frozen features + Attention-based MIL (ABMIL) fine-tuned on 5,000 tiles/WSI via 5-fold CV + 5-model ensembling.	SOTA on 8 biomarker tasks (MSI, HER2, ER, etc.), +1.75 AUC vs. single-shot retraining, statistically on-par with proprietary FMs.	128 V100
RudolfV[117]	The dataset consists of 133,998 slides comprising 34,103 cases, with 1.25 billion image patches extracted for training.	Evaluated using balanced accuracy (classification), F1 scores (per cell type), and similarity scores (reference case search).	Trained with DINOv2, stain color augmentation; fine-tuned by optimizing linear classifier on frozen encoder, sometimes adapting foundation encoder.	Benchmarked on H&E/IHC cell classification, H&E tissue segmentation, IHC biomarker scoring, reference case search, and histological/molecular prediction	16 A100
HIBOU-B[118]	Hibou-B was pre-trained	Hibou-B evaluated with	Trained self-supervised with	Strong performance, achieving	8 A100

	on 512 million clean tissue patches sampled from a proprietary dataset of over 1 million WSIs.	top-1 accuracy (patch-level, linear probe) and AUC (slide-level, attention pooling).	DINOv2 + registers on image patches; downstream: frozen feature extractor, trained linear classifier (patch-level) or attention pooling (slide-level).	competitive results on patch-level benchmarks and surpassing Prov-GigaPath on two out of three slide-level benchmarks despite having significantly fewer parameters.	
HIBOU-L[118]b	Hibou-L was pre-trained on 1.2 billion clean tissue patches derived from a proprietary dataset of over 1 million WSIs.	top-1 accuracy for patch-level tasks and AUC for slide-level classification using an attention-based pooling model.	Trained self-supervisedly using DINOv2, fine-tuned by keeping the feature extractor frozen while training a linear classifier or attention pooling model.	Hibou-L achieved state-of-the-art performance, attaining the highest average accuracy across six patch-level datasets and the highest AUC on all three slide-level benchmarks.	32 A100
H-Optimus-0[c]	Over 500,000 H&E-stained WSIs, from which tiles were extracted, with an average of 1.5 slides per patient.	Tile-level: mean accuracy over 3 runs; slide-level: average AUC-ROC over 50 ABMIL trainings.	Trained self-supervised; downstream: linear classifier (tile-level) or ABMIL (slide-level) on frozen features, no backbone fine-tuning.	Achieves SOTA performance on tile- and slide-level tasks; outperforms GigaPath, Hibou-B, UNI on multiple benchmarks, especially in CRC and breast cancer detection.	8 A100
Madeleine[119]	The Acrobat dataset with 4,211 whole slide images (WSIs) from 1,153 breast cancer cases, and the BWH Kidney dataset with 12,070 WSIs from 1,069 renal transplant cases.	The evaluation metrics used for this model include macro-AUC for classification tasks, concordance-index (c-index) for survival prediction, and accuracy for IHC quantification.	Trained with global InfoNCE (slide-level) and local GOT (patch-level) losses; fine-tuned via linear probing, prototyping, or full fine-tuning on morphological/molecular subtyping, survival prediction, IHC quantification.	The model outperforms all baselines in 13/13 few-shot classification tasks, achieving significant gains such as +10.1% over the Intra baseline in TCGA Breast. In survival prediction on TCGA Breast, it reaches a concordance-index of 0.71, outperforming all baselines.	3 RTX3090Ti
COBRA[120]	Trained on 3048 WSIs	The model's performance	The model was trained using a	The model achieved at least +4.4%	4 A100

	(2848 patients) from 5 TCGA cohorts (BRCA, CRC, LUAD, LUSC, STAD); validated on 1604 WSIs (444 patients) from 4 CPTAC cohorts.	was evaluated using the area under the receiver operating characteristic (AUC) for downstream classification tasks.	contrastive loss function with a batch size of 1024 across four NVIDIA A100 GPUs for 2000 epochs. It was not fine-tuned on downstream tasks but evaluated in a self-supervised manner.	AUC improvement over state-of-the-art slide encoders on four different public CPTAC cohorts, despite being pretrained on only 3048 WSIs.	
PLUTO[121]	Pre-trained on 195M image tiles from 158,852 WSIs across 50+ sources; covers 16 tissue groups, 28 disease areas.	Metrics: macro-F1, AUROC; accuracy, balanced accuracy (tile classification); Dice, IoU, bPQ, mPQ, AJI	PLUTO trained with modified DINOv2 (MAE, Fourier loss) on diverse dataset; FlexiViT-S backbone; fine-tuned with task-specific heads, frozen backbone.	Superior out-of-distribution results on NSCLC subtyping and HER2 scoring; matches or outperforms on CRC-100K, Camelyon17-WILDS, GlaS, PanNuke	64 A40
HIPT[122]	Pretrained on 104M 256×256 image patches and 408,218 4096×4096 regions from 10,678 WSIs across 33 cancer types in TCGA.	The model was evaluated using 10-fold cross-validated AUC for slide-level classification tasks and the c-Index (concordance index) for survival prediction.	Self-supervised pretraining with DINO on ViT256-16 and ViT4096-256, followed by fine-tuning of the lightweight ViTWSI-4096 module using Adam optimizer with gradient accumulation.	Superior performance compared to state-of-the-art weakly-supervised methods (e.g., ABMIL, CLAM-SB, GCN-MIL) in both slide-level classification and survival prediction, particularly excelling in low-data regimes and context-aware tasks.	NA
PathoDuet[123]	Pretrained on ~11,000 H&E WSIs from TCGA; fine-tuned on 21,126 paired H&E-IHC patches from HyReCo and BCI datasets.	Evaluated using Acc, AUC, F1-score across downstream classification and prediction tasks.	Trained via self-supervised learning; no labels for pretraining; fine-tuned on downstream tasks with limited labeled data.	Achieves SOTA or competitive results; outperforms ImageNet-pretrained, CTransPath, UNI in patch classification and WSI-level diagnosis.	8 A100, 4 RTX3090
Kaiko[124]	Trained on 29K TCGA	Evaluated with top-1 acc,	Self-supervised pretraining	Achieves competitive or SOTA on	16 H100

	WSIs (FF & FFPE)	balanced acc, DICE (segmentation), ODCorr, RankMe (representation).	(DINO/DINOv2) on online-extracted patches; fine-tuned via linear probe or full fine-tuning.	BACH, CRC, MHIST, PCam, CoNSeP; outperforms Phikon, Lunit.	
PathOrchestra [125]	Trained on 300K high-quality WSIs from three major centers; diverse populations and tissue types.	Evaluated on 112 tasks using AUC, ACC, F1, Dice, mAP.	Pre-trained with DINOv2 (DINO + iBOT) on unlabeled WSIs; fine-tuned via ABMIL or linear probing.	Achieves >0.930 AUC in pan-cancer classification; outperforms GigaPath, UNI in gene expression; >0.950 ACC/F1 in lymphoma, CRC detection.	32 A100
THREADS[126]	Pre-trained on 47,171 WSIs (TCGA, GTEx, MGH, BWH); evaluated across tasks with 93–1,900+ patient cohorts.	Evaluated by AUC (classification) and C-index (survival prediction).	Trained CLIP-style with contrastive loss; cosine-decayed LR after warmup; downstream: gene encoder fully fine-tuned.	Outperforms GigaPath, PRISM, CHIEF on 54 tasks; superior in cancer subtyping, mutation prediction, treatment response, survival.	4 A100
H0-mini[127]	Distilled on 43M tiles from 6,093 TCGA slides (16 cancers); evaluated on EVA, HEST, and private PLISM, BreastBm cohorts.	Evaluated with balanced accuracy, MonaiDiceScore, Pearson correlation; robustness via cosine similarity, top-k accuracy across scanning variations.	H0-mini trained via knowledge distillation from H-Optimus-0 using DINO+iBOT on unlabeled tiles; downstream: mean-pooling logistic regression on frozen features, no fine-tuning.	Achieves competitive performance on EVA/HEST; matches or surpasses larger models in patch/slide-level and gene expression tasks.	128 V100
TissueConcepts[128]	Trained on ~912K patches from ~7,042 WSIs across 14 public/private datasets.	Evaluated using accuracy, weighted F1, AUC on classification and segmentation tasks.	End-to-end supervised multi-task learning on 16 tasks; fine-tuned by freezing encoder and training task-specific head.	Achieves competitive vs. self-supervised and ImageNet-pretrained models; excels in cross-center generalization and data efficiency.	RTX A5000
OmniScreen[12	Trained on 51,747 WSIs	Assessed using AUC, AP,	Uses Virchow2 embeddings;	Across 1,637 biomarkers/43 cancers,	16 A100

[9]	(41,468 patients); tuned on 10,710 WSIs; evaluated on 10,645 MSK + 2,281 TCGA WSIs.	sensitivity, specificity, PPV, NPV for slide/sample-level predictions.	trains feed-forward aggregator with attention for 50 epochs; selects checkpoints maximizing mean AUC/AP; sets per-label thresholds at 90% sensitivity.	mean AUC 0.84 (sensitivity 0.92, specificity 0.55) on MSK samples; comparable TCGA results; top markers ≥ AUC 0.90; rare mutation screening achieves 100% NPV.	
BROW[130]	Pretrained on >11K WSIs and 180M+ patches from diverse organs/stains; evaluated on 10+ datasets including.	Evaluated with ACC, AUC (classification); DICE, AJI, DQ, SQ, PQ (nuclei segmentation).	Self-supervised training via self-distillation with multi-scale views, color aug, patch shuffling, MIM; adapts to downstream tasks with minimal/no fine-tuning.	Achieves SOTA/competitive results: significant gains over ImageNet models in slide subtyping, 97.83% ACC on SIPaKMeD, strong CoNSeP metrics; demonstrates robustness and generalization.	10 A100
BEPH[131]	Pre-trained on 11.77M histopathology tiles; comparable to earlier studies but smaller than current SOTA models (e.g., Prov-GigaPath48).	Evaluated using AUC (classification), C-index and p-values (survival), mean attention scores (heatmap analysis).	Trained via self-supervised MIM on unlabeled tiles; fine-tuned on labeled downstream tasks; online fine-tuning available via web platform.	Strong performance: outperforms in survival prediction, competitive in WSI classification; attention effectively highlights cancerous vs. non-cancerous regions.	8 A100
Atlas[132]	Trained on 1.2M WSIs (490K+ cases), generating 3.4B tiles; evaluated on 21 public benchmarks.	Evaluated using balanced accuracy (classification) and Pearson correlation (regression) on morphology- and molecular-related tasks.	Self-supervised training with adapted RudolfV (DINOv2-based); fine-tuned via linear probing with frozen backbone.	Achieved SOTA average performance (61.9% across 21 benchmarks), outperforming Virchow2, H-Optimus-0; top results in BACH (93.1%), CRC-100k (97.1%), MSI STAD (76.0%).	H100

[a]: Reference numbers here indicated reference in the main text.

Abbreviations: WSIs: Whole-slide images; Paras: Parameters; ViT: Vision Transformer; TCGA: The Cancer Genome Atlas; TULIP: The University of Louisville

Image Processing dataset; MoCo: Momentum Contrast; SwAV: Swapping Assignments between Views; DINO: Self-Distillation with No Labels; mPQ: mean Panoptic Quality; ACC: Accuracy; AUC: Area Under the ROC Curve; F1: F1 score; SRCL: Semantic-aware Representation learning with Contrastive Learning; SVM: Support Vector Machine; iBOT: masked image modeling with inpainting and distillation; MIL: Multiple Instance Learning; ROC: Receiver Operating Characteristic; C-index: Concordance index; ID: In-Distribution; OOD: Out-of-Distribution; MSKCC: Memorial Sloan Kettering Cancer Center; DeLong's test: Statistical test for comparing ROC curves; ViT: Vision Transformer; AP: Average Precision; κ: Cohen's Kappa; ISUP: International Society of Urological Pathology; ABMIL: Attention-based Multiple Instance Learning; OT: Organoid and Tumor; PRAD: Prostate Adenocarcinoma; H-optimus-0: A high-performance histopathology foundation model; GOT: Global-Local Optimization with Triplet loss; CPTAC: Clinical Proteomic Tumor Analysis Consortium; MAE: Masked Autoencoder; Fourier loss: Loss function based on Fourier transform for texture consistency; FlexiViT: Flexible Vision Transformer; NSCLC: Non-Small Cell Lung Cancer; HER2: Human Epidermal Growth Factor Receptor 2; CRC-100K: Colorectal Cancer 100,000 dataset; GlaS: Gland Segmentation Challenge; PanNuke: Pan-cancer Nuclei Segmentation dataset; HyReCo: Hybrid Registration and Co-localization dataset; BCI: Breast Cancer Immunohistochemistry dataset; FF: Formalin-Fixed; FFPE: Formalin-Fixed Paraffin-Embedded; ODCorr: Object Detection Correlation; RankMe: Metric for evaluating representation quality; GTEx: Genotype-Tissue Expression project; MGH: Massachusetts General Hospital; BWH: Brigham and Women's Hospital; EVA: Evaluation of Vision models in pathology; HEST: Histopathology Embedding Similarity Test; PLISM: Private Lymphoma Subtyping and Mutation dataset; MonaiDiceScore: Dice score implementation from MONAI library; PPV: Positive Predictive Value; NPV: Negative Predictive Value; AJI: Aggregated Jaccard Index; DQ: Detection Quality; SQ: Segmentation Quality; PQ: Panoptic Quality; MIM: Masked Image Modeling; BEPH: Biomedical Entity-aware Pathology Foundation Model.

Supplementary Table 7

Supplementary Table 7 Methodological details of multimodal large language models.

Model	Dataset sizes	Evaluation metrics	Fine-tuning strategies	Performance benchmarks	GPU
PLIP[136]a	Trained on 1.1M image-text pairs (Twitter) + 2.3M (PathLAION); validated on Kather, PanNuke, DigestPath, WSSS4LUAD, PathPedia, PubMed, Books.	Evaluated using Recall@10, Recall@50, weighted/macro F1, MCC for zero-shot classification and text-to-image retrieval.	Pre-trained via contrastive learning on image-text pairs; fine-tuned with linear probing and 1–100% data on Kather, etc.	Outperforms baselines: Recall@10=0.409, Recall@50=0.752 on PathPedia; superior zero-shot accuracy vs. CLIP and MuDiPath.	NA
HistGen[137]	Trained and evaluated on ~7,800 TCGA WSI-report pairs; external datasets (UBC-OCEAN, Camelyon, TUPAC16, six TCGA cohorts) for transfer learning.	Assessed using BLEU, METEOR, ROUGE-L (reporting); accuracy, AUC (subtyping); c-Index (survival).	End-to-end pretraining on WSI-report pairs; downstream fine-tuning via region-to-WSI pooling with classifier/survival head.	HistGen outperforms SOTA in WSI report generation and excels in subtyping and survival analysis, showing strong transferability.	NA
PathAlign[138]	Trained on 354,089 WSI-text pairs (DS1) + 12,268 TCGA WSIs with synthesized text.	Evaluated using MAP, NDCG, ROUGE-L, METEOR, clinician-rated accuracy, AUC, and balanced accuracy.	Stage 1: trained with ITC and ITM losses for image-text alignment; Stage 2: fine-tuned with frozen PaLM-2 S for text generation/retrieval.	Achieved 0.945 AUC (NSCLC), 0.971 AUC (RCC), and 73.5% top-1 retrieval accuracy, showing strong clinical and retrieval performance.	NA
CHIEF[139]	Validated on 13,661 WSIs from 15 datasets (11 cancers) across global hospitals.	Primary metric: AUROC with 95% CIs (bootstrapping); significance tested via two-sided Wilcoxon signed-rank test.	Trained with weak supervision; unsupervised tile pre-training + weakly supervised WSI pre-training; fine-tuned on downstream tasks.	CHIEF outperforms SOTA methods (CLAM, ABMIL, DSMIL) across 15 datasets; up to AUROC 0.9943, P = 0.000061.	8 V100

PathGen[140]	PathGen-1.6M: 1.6M image-text pairs from 7,300 WSIs (27 tissue types); PathGeninit: combines 700K pairs from PathCap, Quilt-1M, OpenPath.	Evaluated on zero-shot/few-shot accuracy, WSI classification (F1, AUC), and PathMMU benchmark accuracy for multimodal understanding.	PathGen-CLIP: pre-trained on PathGen-1.6M, fine-tuned on PathGeninit; PathGen-LLaVA: two-stage LLaVA-style alignment of PathGen-CLIP-Linit with Vicuna.	PathGen-CLIP-L achieves 79.7% avg zero-shot accuracy (9 datasets) and 92.6% AUC in WSI classification; PathGen-LLaVA reaches 58.4% on PathMMU, surpassing GPT-4V (49.8%).	NA
PathChat[141]	Trained on 456,916 instructions and 999,202 QA turns from sources like image captions, articles, WSI regions.	Evaluated with accuracy on PathQABench (Private/Public) for multiple-choice diagnostics; human expert preference studies for open-ended responses.	Pretrained pathology image encoder, connected to a 13B-parameter Llama 2 model via multimodal projector; entire system fine-tuned on pathology instruction dataset.	PathChat achieves SOTA in multiple-choice diagnostics across tissue types; outperforms GPT-4V in human evaluations for open-ended diagnostic reasoning and image description tasks.	8 A100 ; 2 RTX 3090
PathAsst[142]	PathCap holds ≈207 k pathology image-caption pairs, PathInstruct adds ≈180 k instruction samples, and PubMed abstracts supply 5.3 M entries for retrieval.	zero-shot F1 for classification, R@k for cross-modal image retrieval, and accuracy/F1 for closed/open questions on PathVQA.	first freeze vision and LLM, train only the FC adapter on detailed descriptions, then unfreeze the LLM and continue instruction-tuning on 35 k curated samples via next-token prediction.	PathCLIP beats PLIP and OpenAI CLIP (up to 11× higher R@10, +1–11 % F1), and PathAsst tops prior MLLMs on PathVQA (≈91 % closed, 38 % open).	NA
Prov-GigaPath[143]	Pre-trained on >170,000 WSIs (Providence, TCGA); fine-tuned/evaluated on TCGA-LUAD and Zenodo subset.	The model's performance was evaluated using AUROC and AUPRC through 10-fold cross-validation across all tasks.	Pre-trained with DINOv2 (tile-level) and masked autoencoder + LongNet (slide-level); fine-tuned with task-specific learning rates, weight decay, and gradient accumulation over 20 epochs.	Prov-GigaPath achieves SOTA in pan-cancer subtyping and mutation prediction, outperforming MI-Zero, BiomedCLIP, PLIP in AUROC and AUPRC.	16 nodes 4 A100

TITAN [144]	Pre-trained on 335,645 WSIs and 423,122 synthetic region-level captions from clinical reports.	Evaluated using balanced accuracy (multi-class), AUROC (binary), quadratic κ (grading/IHC), and c-index (survival).	Trained with self-supervised masked image modeling and vision-language alignment on WSI crops paired with synthetic captions; evaluated off-the-shelf without fine-tuning.	Outperforms PRISM and CHIEF: +8.4% avg gain in balanced accuracy/AUROC (subtyping), +4% in κ (grading), +1.7–3.7% in AUROC (molecular classification).	8 A100
CONCH [145]	Trained on 1.17M paired image-text data (EDU, PMC OA) with unimodal pretraining on 16M unlabeled tiles and 950K+ pathology reports/abstracts.	Evaluated using balanced accuracy, weighted F1, quadratic κ (grading), and Recall@K (retrieval) across classification, grading, and retrieval tasks.	Initialized with domain-specific unimodal pretraining (iBOT for images, autoregressive LM for text); trained via contrastive learning and captioning.	Outperforms PLIP, BiomedCLIP, CLIP across 14 benchmarks; excels in zero-shot classification, retrieval, segmentation, and generalizes well to rare diseases.	8 A100
SlideChat [146]	SlideInstruction: 4,181 WSI-caption and 175,753 VQA pairs from TCGA; SlideBench-Caption (734), SlideBench-VQA(TCGA) (7,827), SlideBench-VQA(BCNB) (7,247), and external WSI-VQA.	Evaluated using BLEU-1/2/3/4, ROUGE-L, GPT-score (captioning) and accuracy (VQA).	Two-stage training: cross-domain alignment of slide encoder and projector (LLM frozen) on 4.2k captions; full fine-tuning of encoder, projector, and Qwen2.5-7B-Instruct.	Surpasses all baselines, achieves SOTA on 18/22 tasks, including 81.17% accuracy on SlideBench-VQA(TCGA) and 54.14% on SlideBench-VQA(BCNB).	NA
PMPRG [147]	The model was trained and evaluated on a dataset of 7,422 whole-slide images collected from 1,991 patients across colon and kidney specimens.	Performance was assessed using NLG metrics (BLEU-n, METEOR, ROUGE-L) and clinical-efficacy metrics (F1-score and accuracy over 26 diagnostic classes).	Unsupervised pretraining with DINO on region-level patches; report generation via fine-tuning a frozen PubMed-GPT-2 using tag-guided cross-attention on visual features.	Diagnosis-type accuracy of 0.8485 and tumor-grade accuracy of 0.4242; PMPRG attained METEOR 0.6834, ROUGE-L 0.6033, and clinical-accuracy 0.6022.	NA

MuMo [148]	Trained and evaluated on a multi-center dataset of 429 HER2-positive gastric cancer patients (310 anti-HER2 therapy, 119 anti-HER2 + immunotherapy), with an external cohort of 39 patients.	Primary evaluation: AUC for response prediction; PFS and OS analyzed via log-rank test for risk stratification.	MuMo trained end-to-end on multi-modal data using contrastive learning for feature alignment and learnable embeddings to handle missing modalities; no prior fine-tuning from existing models.	Achieved AUC 0.821 (anti-HER2) and 0.914 (combination therapy), outperforming single-modality models; low-risk group identified by MuMo showed significantly prolonged PFS and OS.	NA
ConcepPath[149]	The model was evaluated on three public TCGA datasets: NSCLC with 1042 lung cancer cases, BRCA with 933 breast cancer cases, and STAD with 268 gastric cancer cases.	The model's performance was assessed using the area under the receiver operating characteristic curve (AUC) and accuracy (ACC).	Leverages a pre-trained CLIP-based pathology vision-language model as a fixed feature extractor and concept aligner; trained within a Multiple Instance Learning framework to align image patches with GPT-4-induced expert concepts and learnable data-driven concepts.	The model significantly outperformed seven state-of-the-art methods on all five evaluated tasks, achieving notable improvements, such as a nearly 7% increase in AUC for classifying EBV-positive gastric cancer cases.	8 RTX 3090
GPT-4V[150]	Evaluated on CRC100K (100K tiles), MHIST (1,050 images), and PatchCamelyon (PCAM, 96K patches), with only small subsets used for in-context learning.	Primary metric: accuracy, with 95% CIs from 100,000 bootstrap iterations; supplemented by per-class recall and confusion matrices.	GPT-4V was not trained or fine-tuned; used in-context learning (ICL) with few-shot examples in the prompt, no parameter updates or backpropagation.	Accuracy: 90% (10-shot) on CRC100K, 83.4% on MHIST, 88.3% on PCAM (10-shot kNN-sampled), narrowing the gap with fine-tuned foundation models like Phikon and UNI.	NA
MINIM[151]	Trained on large-scale multimodal medical data: 11,438 paired OCT images/reports, 3,745	Evaluated using FID, Inception Score (IS), MS-SSIM, CAS, IIR, ITR, and clinical expert	Trained by integrating image-text pairs into a stable diffusion framework; fine-tuned with self-	Achieves SOTA performance: FID 65.3 (OCT), 32.7 (fundus), 110.4 (X-ray), 94.8 (CT); CAS	8 A100

	chest CT/X-rays, 1,960 brain MRIs, and 3,694 breast MRIs, with textual descriptions.	subjective scoring.	improvement strategies including RLHF and transfer learning for new modalities.	79.09% (OCT), 86.16% (fundus), 79.42% (CT), 77.23% (X-ray), outperforming existing methods.	
PathM 3[152]	PatchGastric dataset includes 991 whole-slide images divided into 20% training, 40% validation, and 40% testing subsets.	Performance metrics include classification accuracy and captioning evaluated with BLEU@4, METEOR, and SPICE scores.	Trained using multi-task joint learning, employing AdamW optimizer and a mixture of cross-entropy and generative loss; the newly introduced query-based transformer and correlation module are fine-tuned.	Achieves 86.40 % classification accuracy and captioning scores of 0.520 BLEU@4, 0.394 METEOR, and 0.591 SPICE, outperforming prior MIL and multimodal baselines.	16 GPU
FGCR[153]	The study used an in-house gastric dataset of 3,598 paired WSIs and reports, and the public GastricADC dataset of 991 slides.	Retrieval performance was measured by MAP@k, P@k, PTACC, and PIoU@k.	Trained end-to-end for 150 epochs with Adam optimizer and cosine scheduler, using ResNet50 patch features pre-trained with BYOL and BERT embeddings.	The proposed model achieved state-of-the-art results on both datasets, outperforming ViT, RetCCL, SISH, DTN, FLIP, WCAP, TransMIL, and SETMIL.	RTX 3090
Prompt Bio[154]	TCGA dataset comprises 482 WSIs and CPTAC dataset comprises 105 WSIs.	Area under the ROC curve (AUC) is reported on MSI and BRAF classification tasks.	End-to-end trained with Adam optimizer, binary cross-entropy loss; PLIP encoders frozen, only task-specific layers updated.	91.49% AUC (TCGA MSI), 80.25% AUC (TCGA BRAF), and 91.25% AUC (CPTAC MSI).	RTX 3090
HistoC ap[155]	The study employed 25 120 whole-slide images (23 517 training, 603 validation, 1 000 testing) from GTEx for downstream captioning.	Performance was assessed with tissue-type classification accuracy, BLEU-4, ROUGE-L and METEOR scores for generated captions.	After freezing the DINO-pre-trained HIPT vision encoder, the model was end-to-end fine-tuned using the Adam optimizerwith only the attention layer.	The best configuration achieved 89.53 % tissue-type accuracy, BLEU-4 = 0.578, ROUGE-L = 0.742 and METEOR = 0.703 on the held-out 1 000-image test set.	A100

mSTAR[156]	Pretrained on 26,169 slide-level modality pairs (8,440 WSI–Report, 8,965 WSI–RNA-Seq, 8,764 Report–RNA-Seq) from 10,275 patients (32 cancers), totaling >116M patches; evaluated on 97 downstream tasks across public (TCGA, CAMELYON, PANDA) and private (NFH, ZJ1) cohorts.	Metrics: Macro-AUC (95% CI) for classification, C-Index (95% CI) for survival, Recall@K for zero-shot retrieval, and BLEU/METEOR/ROUGE-L for report generation.	Contrastive pretraining of a slide-level aggregator on multimodal pairs; Self-teaching of a ViT-L patch encoder by matching patch features to those from the frozen aggregator. Downstream tasks freeze the ViT-L extractor and train ABMIL or multimodal heads from scratch using Adam/cosine schedules.	Sets new SOTA across 15 categories: avg Macro-AUC 0.845 (diagnosis), 0.697 (mutation), 0.763 (IHC), 0.886 (subtyping); C-Index 0.686 (OS), 0.708 (multimodal fusion); leads in zero-shot classification, retrieval, and report generation.	4 H800
GPT-4 Enhanced[157]	Evaluated on 20 simulated multimodal patient cases featuring clinical notes, imaging reports, pathology images, and genomic data for GI cancers.	Performance assessed using tool-use accuracy, clinical conclusion correctness, proportion of sub-questions addressed, and citation accuracy.	Reasoning agent using pre-trained GPT-4 without fine-tuning; autonomously selects and applies external tools based on input context.	Achieved 91.0% clinical conclusion accuracy and 87.5% tool-use accuracy, significantly outperforming standalone GPT-4 (30.3% accuracy).	NA
PRISM[158]	Pre-trained on a large, diverse dataset of whole slide images (WSIs) paired with clinical reports; evaluated on TCGA and MSKCC datasets, with biomarker prediction tasks using subsets (10%–100% data) for label efficiency studies.	Performance measured using AUC (binary biomarker/cancer detection), accuracy, and comparisons to supervised baselines for subtyping and detection.	Pre-trained generatively to produce clinical reports using a Perceiver-based slide encoder and BioGPT-based language decoder; then fine-tuned by adapting the pre-trained slide encoder on smaller datasets for detection, subtyping, and prediction.	Zero-shot classification surpasses supervised baselines; fine-tuning outperforms training from scratch and zero-shot methods in subtyping; excels in low-data regimes for biomarker prediction, often exceeding supervised models trained on full datasets.	16 V100 32GB
HistoGPT[159]	Trained on 15,129 WSIs from 6,705 patients; evaluated on	Performance assessed using: semantic similarity (BioBERT,),	Two-stage training: (1) image-only pre-training via multiple	Outperforms BioGPT-1B and GPT-4V in text accuracy and	A100 -SX

	internal and external cohorts including Munich, Münster, Mayo Clinic, and Radboud University Medical Center datasets.	keyword accuracy (Dictionary, ScispaCy Jaccard), NLP scores (BLEU-4, ROUGE-L, METEOR, BERTscore), clinical expert ratings (blinded study), and task-specific metrics (RMSE).	instance learning (MIL) on WSIs; (2) image-text fine-tuning using autoregressive language modeling on paired WSI-report data.	similarity; 45% of generated reports rated indistinguishable from expert reports; strong zero-shot performance with RMSE=1.8mm (tumor thickness) and F1=0.63 (BCC subtyping).	M4
PathologyVLM [160]	Trained on a curated pathology dataset: 827,401 image-text pairs (PCaption-0.8M), 518,413 refined alignment pairs (PCaption-0.5M), and 35,543 VQA instruction pairs from PathVQA and PMC-VQA.	Evaluated using accuracy (closed-set) and recall (open-ended) for supervised VQA; accuracy, recall, and precision for zero-shot classification tasks.	Pretrain PLIP model on image-caption data; domain alignment via training connector and LoRA modules for image description generation, with encoder and LLM frozen; end-to-end VQA instruction fine-tuning on VQA dataset.	Achieves SORT performance among similarly-sized multimodal models on both supervised (PathVQA, PMC-VQA) and zero-shot VQA benchmarks, outperforming general-purpose LLaVA and medical VLMs such as LLaVA-Med and Quilt-LLaVA.	16 A100
MUSK [161]	Trained on a massive scale: 50 million pathology image patches from ~33,000 whole-slide images and 1 billion pathology-related text tokens for masked pretraining, followed by contrastive learning on 1 million high-quality image-text pairs from QUILT-1M and PathAsst.	Evaluated across diverse tasks using balanced accuracy (linear probe, few-shot, zero-shot classification), VQA accuracy, Recall@50 (I2T and T2I retrieval), and mMV@5 (I2I retrieval).	Masked modeling on unpaired images and text to learn general representations; contrastive learning on curated image-text pairs to align vision and language modalities. Downstream tasks require little or no fine-tuning, leveraging the robust pretraining foundation.	Achieves state-of-the-art performance, significantly outperforming models like PLIP, QUILT-1M, BiomedCLIP, and CONCH across 23 pathology benchmarks, including zero-shot classification, cross-modal retrieval, and visual question answering.	64 V100

[a]: Reference numbers here indicated reference in the main text.

Abbreviations: WSIs: Whole-slide images; ITM: Image-text matching; ITC: Image-text contrastive; NLG: Natural language generation; PFS: Progression-free survival; OS: Overall survival; RMSE: Root mean square error; AUC: Area under the receiver operating characteristic curve; AUROC: Area under the ROC curve; AUPRC: Area under the precision-recall curve; F1: F1 score; MCC: Matthews correlation coefficient; MAP: Mean average precision; NDCG: Normalized discounted cumulative gain; ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation (Longest common subsequence); BLEU: Bilingual evaluation understudy; METEOR: Metric for Evaluation of Translation with Explicit ORdering; SPICE: Semantic Propositional Image Caption Evaluation; C-Index: Concordance index; κ: Cohen's kappa; PTACC: Patch-to-tile accuracy; PIoU: Patch IoU; CAS: Clinical acceptability score; FID: Fréchet inception distance; IS: Inception score; MS-SSIM: Multi-scale structural similarity; IIR: Image-to-image retrieval; ITR: Image-to-text retrieval; kNN: k-Nearest neighbors; CI: Confidence interval; LoRA: Low-rank adaptation; VQA: Visual question answering; I2T: Image-to-text; T2I: Text-to-image; I2I: Image-to-image; mMV@5: Mean multi-view recall at 5; NLG: Natural language generation; RMSE: Root mean square error.

Supplementary Table 8

Supplementary Table 8 Description of related Artificial Intelligence Term.

Term	Description
Transformer architectures	A type of deep learning model designed to handle sequential data without relying on RNNs, featuring self-attention mechanisms.
Self-supervised learning (SSL)	A method for training models using input data alone, without labeled responses, to learn useful features or representations.
Zero- or few-shot transfer learning	Techniques that enable models to make predictions for tasks with little or no prior examples by leveraging knowledge learned from other tasks.
Masked Language Modeling (MLM)	A pre-training objective where parts of the input are masked and the model learns to predict these masked tokens based on context.
Contrastive learning	A method in SSL where the model learns to distinguish between similar and dissimilar data pairs, improving its understanding of data representations.
Pre-training and fine-tuning paradigm	A two-step process where models are first trained on a large dataset and then adjusted on smaller datasets tailored to specific tasks.
Self-attention mechanism	A component in transformers that allows each position in the sequence to attend to all positions in the previous layer, capturing dependencies regardless of their distance.
Generative adversarial networks (GANs)	A framework for training generative models through an adversarial process, involving two neural networks contesting with each other.
Diffusion models	A class of generative models that learn to reverse a diffusion process, effectively generating new data samples.
Mixture of Experts (MoE) architecture	A model design that uses a gating network to select which parts of the model ("experts") should be activated for a given input.
Parameter-efficient fine-tuning	Methods that adjust only a small portion of a model's parameters during fine-tuning, reducing computational costs while maintaining performance.

Low-Rank Adaptation (LoRA)	A technique for efficient fine-tuning that modifies low-rank matrices instead of full layers, significantly lowering the number of trainable parameters.
Adapter-based fine-tuning	An approach to fine-tuning where small "adapter" modules are inserted into a pre-trained model, allowing it to adapt to new tasks with minimal changes.
Black-box problem	The challenge of understanding how a machine learning model makes decisions, especially when its internal workings are opaque.
Knowledge distillation	A technique where a smaller model is trained to mimic the behavior of a larger, more complex model, often leading to faster inference times.
Prompt engineering	Crafting specific prompts or instructions to guide language models to produce desired outputs, enhancing their applicability to various tasks.
Transfer learning	Leveraging knowledge gained from solving one problem to help solve a different but related problem.
Multimodal integration	Combining data from different sources (e.g., images, text) to enhance the performance of AI systems.
Model-as-a-Service (MaaS)	A cloud-based service that provides access to pre-trained models, enabling users to integrate AI capabilities into their applications without needing extensive AI expertise.

Reference

1 Harkema, H. *et al.* Developing a natural language processing application for measuring the quality of colonoscopy procedures. *J Am Med Inform Assoc* **18 Suppl 1**, i150-156, doi:10.1136/amiajnl-2011-000431 (2011).

2 Denny, J. C. *et al.* Natural language processing improves identification of colorectal cancer testing in the electronic medical record. *Med Decis Making* **32**, 188-197, doi:10.1177/0272989X11400418 (2012).

3 Mehrotra, A. *et al.* Applying a natural language processing tool to electronic health records to assess performance on colonoscopy quality measures. *Gastrointest Endosc* **75**, 1233-1239 e1214, doi:10.1016/j.gie.2012.01.045 (2012).

4 Wagholikar, K. *et al.* in *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology.* 12-21.

5 Imler, T. D., Morea, J., Kahi, C. & Imperiale, T. F. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. *Clin Gastroenterol Hepatol* **11**, 689-694, doi:10.1016/j.cgh.2012.11.035 (2013).

6 Imler, T. D., Morea, J. & Imperiale, T. F. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. *Clin Gastroenterol Hepatol* **12**, 1130-1136, doi:10.1016/j.cgh.2013.11.025 (2014).

7 Blumenthal, D. M., Singal, G., Mangla, S. S., Macklin, E. A. & Chung, D. C. Predicting Non-Adherence with Outpatient Colonoscopy Using a Novel Electronic Tool that Measures Prior Non-Adherence. *Journal of General Internal Medicine* **30**, 724-731, doi:10.1007/s11606-014-3165-6 (2015).

8 Raju, G. S. *et al.* Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. *Gastrointestinal Endoscopy* **82**, 512-519, doi:10.1016/j.gie.2015.01.049 (2015).

9 Sada, Y., Hou, J., Richardson, P., El-Serag, H. & Davila, J. Validation of Case Finding Algorithms for Hepatocellular Cancer From Administrative Data and Electronic Health Records Using Natural Language Processing. *Med Care* **54**, e9-14, doi:10.1097/MLR.0b013e3182a30373 (2016).

10 Nayor, J., Borges, L. F., Goryachev, S., Gainer, V. S. & Saltzman, J. R. Natural Language Processing Accurately Calculates Adenoma and Sessile Serrated Polyp Detection Rates. *Dig Dis Sci* **63**, 1794-1800, doi:10.1007/s10620-018-5078-4 (2018).

11 Becker, M., Kasper, S., Bockmann, B., Jockel, K. H. & Virchow, I. Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation. *Int J Med Inform* **127**, 141-146, doi:10.1016/j.ijmedinf.2019.04.022

(2019).
12 Denny, J. C. *et al.* Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *J Am Med Inform Assoc* **17**, 383-388, doi:10.1136/jamia.2010.004804 (2010).
13 Parthasarathy, G., Lopez, R., McMichael, J. & Burke, C. A. A natural language-based tool for diagnosis of serrated polyposis syndrome. *Gastrointest Endosc* **92**, 886-890, doi:10.1016/j.gie.2020.04.077 (2020).
14 Fevrier, H. B., Liu, L., Herrinton, L. J. & Li, D. A Transparent and Adaptable Method to Extract Colonoscopy and Pathology Data Using Natural Language Processing. *J Med Syst* **44**, 151, doi:10.1007/s10916-020-01604-8 (2020).
15 Li, D., Udaltsova, N., Layefsky, E., Doan, C. & Corley, D. A. Natural Language Processing for the Accurate Identification of Colorectal Cancer Mismatch Repair Status in Lynch Syndrome Screening. *Clin Gastroenterol Hepatol* **19**, 610-612 e611, doi:10.1016/j.cgh.2020.01.040 (2021).
16 Bae, J. H. *et al.* Natural Language Processing for Assessing Quality Indicators in Free-Text Colonoscopy and Pathology Reports: Development and Usability Study. *JMIR Med Inform* **10**, e35257, doi:10.2196/35257 (2022).
17 Song, G. *et al.* Natural Language Processing for Information Extraction of Gastric Diseases and Its Application in Large-Scale Clinical Research. *Journal of Clinical Medicine* **11**, 2967, doi:10.3390/jcm11112967 (2022).
18 Li, J. *et al.* A deep learning and natural language processing-based system for automatic identification and surveillance of high-risk patients undergoing upper endoscopy: A multicenter study. *EClinicalMedicine* **53**, 101704, doi:10.1016/j.eclinm.2022.101704 (2022).
19 Laique, S. N. *et al.* Application of optical character recognition with natural language processing for large-scale quality metric data extraction in colonoscopy reports. *Gastrointestinal Endoscopy* **93**, 750-757, doi:10.1016/j.gie.2020.08.038 (2021).
20 Seong, D., Choi, Y. H., Shin, S.-Y. & Yi, B.-K. Deep learning approach to detection of colonoscopic information from unstructured reports. *BMC Medical Informatics and Decision Making* **23**, doi:10.1186/s12911-023-02121-7 (2023).
21 Tinmouth, J. *et al.* Validation of a natural language processing algorithm to identify adenomas and measure adenoma detection rates across a health system: a population-level study. *Gastrointest Endosc* **97**, 121-129 e121, doi:10.1016/j.gie.2022.07.009 (2023).
22 Nguyen Wenker, T. *et al.* Using Natural Language Processing to Automatically Identify Dysplasia in Pathology Reports for Patients With Barrett's Esophagus. *Clin Gastroenterol Hepatol* **21**, 1198-1204, doi:10.1016/j.cgh.2022.09.005 (2023).

23 Ganguly, E. K. *et al.* An Accurate and Automated Method for Adenoma Detection Rate and Report Card Generation Utilizing Common Electronic Health Records. *J Clin Gastroenterol* **58**, 656-660, doi:10.1097/MCG.0000000000001915 (2024).

24 Benson, R. *et al.* Leveraging Natural Language Processing to Extract Features of Colorectal Polyps From Pathology Reports for Epidemiologic Study. *JCO Clin Cancer Inform* **7**, e2200131, doi:10.1200/CCI.22.00131 (2023).

25 Li, J. *et al.* An interpretable deep learning framework for predicting liver metastases in postoperative colorectal cancer patients using natural language processing and clinical data integration. *Cancer Med* **12**, 19337-19351, doi:10.1002/cam4.6523 (2023).

26 Bertasius, G., Wang, H. & Torresani, L. in *Proceedings of the 38th International Conference on Machine Learning* Vol. 139 (eds Meila Marina & Zhang Tong) 813--824 (PMLR, Proceedings of Machine Learning Research, 2021).

27 Pan, J., Lin, Z., Zhu, X., Shao, J. & Li, H. in *Proceedings of the 36th International Conference on Neural Information Processing Systems* Article 1919 (Curran Associates Inc., New Orleans, LA, USA, 2022).

28 Wang, Q. *et al.* in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021.* (eds Marleen de Bruijne *et al.*) 24-32 (Springer International Publishing).

29 Karaev, N. *et al.* in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXII* 18–35 (Springer-Verlag, Milan, Italy, 2024).

30 Liu, W., Li, Z., Li, C. & Gao, H. ECTransNet: An Automatic Polyp Segmentation Network Based on Multi-scale Edge Complementary. *Journal of Digital Imaging* **36**, 2427-2440, doi:10.1007/s10278-023-00885-y (2023).

31 Dong, B. *et al.* Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. *CAAI Artificial Intelligence Research*, doi:10.26599/air.2023.9150015 (2023).

32 Zheng, X. *et al.* in *Web and Big Data Lecture Notes in Computer Science* Ch. Chapter 2, 18-30 (2024).

33 Yao, H. *et al.* Unsupervised Segmentation of Colonoscopy Images. *CoRR* **abs/2312.12599**, doi:10.48550/ARXIV.2312.12599 (2023).

34 Chaitanya, K. *et al.* in *Machine Learning in Medical Imaging: 15th International Workshop, MLMI 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 6, 2024, Proceedings, Part II* 201–211 (Springer-Verlag, Marrakesh, Morocco, 2024).

35 Oquab, M. *et al.* DINOv2: Learning Robust Visual Features without Supervision. *arXiv e-prints* (2023). <https://arxiv.org/abs/2304.07193>.

36 Liu, S. *et al.* in *Computer Vision – ECCV 2024.* (eds Aleš Leonardis *et al.*) 38-55 (Springer Nature Switzerland).

37 Kirillov, A. *et al.* Segment Anything. *arXiv e-prints* (2023). <https://arxiv.org/abs/2304.02643>.

38 Ma, J. *et al.* Segment anything in medical images. *Nature Communications* **15**, doi:10.1038/s41467-024-44824-z (2024).

39 OpenAI *et al.* GPT-4 Technical Report. *arXiv e-prints* (2023). <https://arxiv.org/abs/2303.08774>.

40 Team, G. *et al.* Gemini: A Family of Highly Capable Multimodal Models. *arXiv e-prints* (2023). <https://arxiv.org/abs/2312.11805>.

41 Radford, A. *et al.* in *Proceedings of the 38th International Conference on Machine Learning.* (eds Meila Marina & Zhang Tong) 8748--8763 (PMLR).

42 Zhang, S. *et al.* A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *NEJM AI* **2**, AIoa2400640, doi:doi:10.1056/AIoa2400640 (2025).

43 Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674-10685, doi:10.1109/cvpr52688.2022.01042 (2022).

44 Awadalla, A. *et al.* OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv e-prints* (2023). <https://arxiv.org/abs/2308.01390>.

45 Moor, M. *et al.* in *Proceedings of the 3rd Machine Learning for Health Symposium* Vol. 225 (eds Hegselmann Stefan *et al.*) 353--367 (PMLR, Proceedings of Machine Learning Research, 2023).

46 Yang, Z. *et al.* The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv e-prints* (2023). <https://arxiv.org/abs/2309.17421>.

47 Cherti, M. *et al.* in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2818-2829.

48 Tsai, Y.-H. H. *et al.* 6558-6569 (Association for Computational Linguistics).

49 Choromanski, K. M. *et al.* in *9th ICLR 2021: Virtual Event, Austria* (2021).

50 Xiong, Y. *et al.* Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**, 14138-14148, doi:10.1609/aaai.v35i16.17664 (2021).

51 Qian, S. & Wang, C. COM: Contrastive Masked-attention model for incomplete multimodal learning. *Neural Networks* **162**, 443-455, doi:10.1016/j.neunet.2023.03.003 (2023).