npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02236-3

Deep multimodal state-space fusion of endoscopic-radiomic and clinical data for survival prediction in colorectal cancer

Ning Wang[1,5], Jiajing Lin[2,5], Wujin Li[2,5], Yahui Lyu[1,5], Yiqing Jiang[3], Zhizhan Ni[4], Qi Huang[4], Hong Chen[2,4] ✉, Qiang Yan[1] ✉ & Chenshen Huang[2,4] ✉

Integrating complementary surface and cross sectional cues is central to preoperative assessment of colorectal cancer, but technically challenging because endoscopic images and pelvic CT encode anatomy at different scales. Here we present HydraMamba, a multimodal selective state space framework that fuses endoscopy and CT for joint lesion segmentation, lesion detection, and survival prediction. The model couples a shared state space backbone with two lightweight modules. Across the endoscopic dataset and the CT dataset, HydraMamba achieved state-of-the-art lesion analysis (endoscopy: Dice 0.856, F1 0.918; CT: Dice 0.812, F1 0.888) and delivered calibrated survival modeling on the CT dataset (Harrell's C index 0.832, Uno's C@1y 0.853, integrated Brier score 0.161, calibration slope ≈1.01). By unifying endoscopic and CT information in a single coherent architecture, HydraMamba provides an accurate and well-calibrated foundation for lesion analysis and prognostication in colorectal cancer.

Colorectal cancer remains a significant health burden, with rectal cancer as a major contributor to morbidity and mortality[1]. Effective management of colorectal tumors relies on accurate lesion characterization and staging, which increasingly benefit from advanced medical imaging and artificial intelligence. In clinical practice, endoscopy (colonoscopy) provides high resolution visualization of mucosal lesions, whereas cross-sectional imaging such as computed tomography (CT) delineates deeper anatomical context and distant spread. These modalities offer complementary information; however, traditionally they have been analyzed in isolation. There is a growing consensus that multimodal learning can unlock synergistic value from heterogeneous data in oncology[2]. For instance, integrating endoscopic and radiologic data has been shown to improve diagnostic accuracy and prognostic modeling. In gastric cancer, a deep learning model that fuses CT scans with endoscopy images achieved an area under the ROC curve (AUC) of 0.93 in predicting pathological stage, significantly outperforming models using either modality alone[3]. Similarly, combining radiologic features with digitized pathology has yielded robust predictors of patient outcomes; a recent multimodal model integrating CT-based and histopathology-based deep features (along with clinical factors) attained high concordance indices (0.74) in survival prediction for head and neck cancer[4–6].

At the same time, deep learning has revolutionized image-based detection and segmentation of lesions in gastroenterology and oncology. Convolutional neural networks and their variants now approach expert level performance in localizing tumors or precancerous polyps on medical images. In colonoscopy, numerous studies have demonstrated automated polyp detection and delineation systems powered by deep learning, which can assist endoscopists by reducing miss rates and providing real-time decision support[7]. On radiologic scans, AI algorithms excel at highlighting tumors and metastases that might be subtle to human observers[8]. These advances are not merely academic—they translate into practical tools that improve diagnostic precision and workflow efficiency. Nevertheless, challenges remain in generalizing these models across diverse patient populations and imaging devices[7]. Robust performance in lesion detection/segmentation requires large, diverse datasets and architectures capable of capturing both fine grained local details and global context.

Beyond diagnosis, there is an urgent need for methods that can leverage imaging data to predict clinical outcomes such as treatment response and survival. Traditional risk stratification in colorectal cancer hinges on clinicopathological factors, but imaging derived biomarkers offer a noninvasive window into tumor biology and patient prognosis. Recent work demonstrates that deep learning models can extract prognostically relevant features

[1]Huzhou Central Hospital, The Fifth School of Clinical Medicine of Zhejiang Chinese Medical University, Huzhou, Zhejiang, China. [2]Fuzhou University Affiliated Provincial Hospital, School of Medicine, Fuzhou University, Fuzhou, Fujian, China. [3]School of Mathematical Sciences, Tongji University, Shanghai, Shanghai, China. [4]School of Medicine, Tongji University, Shanghai, Shanghai, China. [5]These authors contributed equally: Ning Wang, Jiajing Lin, Wujin Li, Yahui Lyu. ✉e-mail: hongruhe@163.com; yianq@hzhospital.com; chenshenhuang@126.com

from medical images that sometimes rival or even surpass classical risk factors in predictive power. In gastric cancer, for example, a CT-based deep learning survival score was shown to stratify patients into high vs. low risk groups with predictive performance comparable to postoperative pathological staging (5-year disease free survival AUC ≈ 0.71)[9]. In bladder cancer, an interpretable deep neural network trained on preoperative CT scans significantly outperformed standard clinical models in forecasting overall survival, and its risk predictions remained independently prognostic in multivariate analysis[10].

Multimodal Learning in Medical Imaging: Harnessing multiple data sources for more informed decision making has been a prominent theme in recent medical AI research. Multimodal learning approaches seek to combine modalities such as medical images, clinical data, and omics in a unified model. In the context of gastrointestinal cancers, researchers have explored fusing endoscopic images with radiological scans to compensate for the limitations of each modality. The study by Zhang et al.[3] on gastric cancer is a representative example: the authors developed an AI model that inputs both CT images and gastroscopy photographs to predict pathological staging. Their integrated model significantly outperformed single modality models, confirming that cross modal synergies can boost diagnostic accuracy. Beyond imaging, integration of pathology and radiology has shown promise. Qian et al. combined deep features from CT scans and whole slide histology images using a Cox proportional hazards model, improving prognostic predictions in head and neck cancer[4]. In colorectal cancer, there is increasing interest in merging colonoscopy findings with imaging and even genomics to refine risk assessment[11]. These works collectively illustrate that multimodal fusion can capture aspects of disease biology that might be missed by unimodal analysis alone. However, designing architectures that effectively align and fuse heterogeneous data remains challenging. Simple late fusion may not fully exploit inter-modal correlations, whereas early fusion (concatenating raw inputs) can be intractable due to differing data dimensions. This has motivated research into attention mechanisms and cross modal transformers that learn shared representations. Our proposed method contributes to this line of work by using a state space model backbone to naturally accommodate sequential endoscopic data alongside volumetric imaging, with an explicit token matching mechanism to align anatomical regions across modalities.

Deep Learning for Lesion Detection and Segmentation: Automated detection and segmentation of lesions in medical images is a well established application of deep learning, with numerous successes reported in recent years. In gastrointestinal endoscopy, AI-driven polyp detection systems have attracted particular attention as they can alert endoscopists to subtle lesions in real time and reduce miss rates. Ali et al. (2024) organized a large-scale colonoscopy computer vision challenge to assess the generalizability of polyp detection/segmentation models[7]. Top-performing algorithms–often based on convolutional neural networks or transformer hybrids achieved high accuracy on the challenge dataset, though the study underscored the need for better robustness across different centers and imaging conditions. The general trend is that deep learning models can now outperform traditional computer aided detection methods by a substantial margin, given sufficient training data. For CT and MRI, organ and tumor segmentation has been revolutionized by 3D CNNs and variants of the U-Net architecture. For example, Chen et al. demonstrated that incorporating masked image modeling pre-training significantly improves 3D tumor segmentation in volumetric scans[8].

Foundation models like the Segment Anything Model (SAM)[12] have also been adapted to medical images, showing encouraging results in zero-shot or few-shot segmentation of various lesion types (as evidenced by recent Nature Communications reports on MedSAM[13]). Nevertheless, challenges such as class imbalance, small lesion size, and boundary ambiguity persist. Our work addresses some of these issues through token anatomy-aware interpolation that emphasizes consistent localization of lesions across modalities, potentially improving both detection and segmentation. By leveraging endoscopy, which directly visualizes mucosal lesions, alongside CT, which provides context for external invasion and lymph nodes, our model aims for more comprehensive lesion delineation in colorectal cancer. This could aid surgeons and oncologists in treatment planning (e.g., identifying all tumor foci preoperatively).

Survival Modeling and Outcome Prediction: Predicting patient outcomes using imaging data has become an important research area, especially with the rise of radiomics and deep learning prognostic models. Conventional radiomics involves handcrafting quantitative features from images and correlating them with outcomes, whereas deep learning can automatically learn complex feature representations optimized for outcome prediction. Several high-profile studies have validated the power of imaging-based predictors. In a study by Nature Communications 2023, Jiang et al. developed a multitask deep learning model for gastric cancer that simultaneously classified the tumor microenvironment (TME) subtype and predicted survival from CT images[9]. In bladder cancer, a recent multicenter study in NPJ Precision Oncology trained a deep neural network on contrast enhanced CT to predict overall survival; the model's performance (validated across four cohorts) demonstrated that deep learning prognostic models can generalize and provide interpretability via techniques like SHAP (highlighting image regions predictive of poor outcome)[10]. These advances suggest that noninvasive imaging, when coupled with artificial intelligence, can uncover subtle prognostic indicators that human observers might overlook. For colorectal cancer, outcome prediction is particularly relevant in the context of neoadjuvant therapy: identifying which patients will achieve a complete response to chemoradiation could spare them unnecessary surgery. Prior works have explored MRI based radiomics for predicting pathological complete response in colorectal cancer, and more recent efforts use deep learning on MRI or CT to improve upon those models. Yet, one limitation in many studies is that they rely on single modalities. Our approach, by integrating endoscopic imagery with CT, has the potential to enhance outcome modeling–since endoscopy might capture tumor phenotype associated with aggressiveness, while CT captures anatomical stage. We incorporate a survival analysis component in our framework (through a Cox proportional hazards layer on top of learned features), and we will demonstrate that the fused feature representations from our multimodal state space model yield superior risk stratification compared to using either modality alone.

State Space and Sequence Models in Vision: The backbone of our proposed architecture aligns with a growing movement in computer vision towards sequence modeling architectures, including transformers and state space models. While convolutional neural networks have long dominated imaging tasks, their local receptive fields and inductive biases can be suboptimal for modeling long-range dependencies or globally contextual information. Transformers[14], with their self-attention mechanism, were introduced to vision (Vision Transformers, ViT[15]) to directly model relationships between distant patches, and quickly proved effective for image recognition and segmentation. However, as noted earlier, transformers incur high computational cost for large images or long sequences, and can be data hungry. Structured state space models (SSMs)[16] emerged as an appealing alternative for certain vision tasks, drawing from sequence modeling successes in NLP. SSMs use linear recurrent layers parameterized by state matrices that can be efficiently computed via convolution, enabling them to handle very long sequences with linear scaling[17]. Our design of a global context injection module is partly inspired by this idea of hybridizing attention with state space kernels. We apply a similar principle not for super resolution, but for the multimodal fusion problem, ensuring our model can grasp global correspondences.

Masked Image Modeling and Self-Supervised Pretraining: Self-supervised learning via masked image modeling (MIM) has become a cornerstone of modern computer vision, following the success of BERT style masked language modeling in NLP. The idea is to mask out parts of the input image and train a model to reconstruct the missing content, thereby forcing the network to learn rich internal representations of the visual data. Methods like MAE (Masked Autoencoder) demonstrated that MIM pretraining can substantially improve downstream task performance by capturing high-level semantic features. Until recently, however, MIM had been

relatively underexplored in the medical imaging domain[8]. This is now rapidly changing. Chen et al. (WACV 2023) showed that MIM pretraining on 3D medical images boosts performance on tasks like tumor segmentation, especially when annotated data are limited[8]. Moreover, researchers have begun tailoring MIM strategies to domain-specific architectures. The work MambaMIM (Tang et al., MedIA 2025) is particularly relevant to us: it was the first attempt to pretrain a Mamba state space model using masked image modeling[18]. MambaMIM introduced a novel token interpolation strategy (TOKI) to maintain consistency across the state dimensions during pretraining, and employed a hierarchical decoder to learn multi-scale features. This approach proved effective, yielding notable improvements on multiple medical segmentation benchmarks[18].

Collectively, the progress in multimodal data integration, image-based lesion analysis, and prognostic modeling sets the stage for next-generation AI systems in colorectal cancer care. However, realizing this vision demands new architectures that can seamlessly handle multimodal, high-dimensional medical data and capture both the spatial context of images and the sequential nature of, say, endoscopic video frames or longitudinal scans. Convolutional networks have dominated medical image analysis to date, but they may struggle to model long-range dependencies or cross-modal relationships explicitly. Transformer-based architectures offer an alternative with global attention, yet vanilla self-attention has limitations: quadratic complexity with respect to sequence length and a restricted input size, making it less ideal for extremely high resolution images or long image sequences[17]. In response, researchers have begun exploring state space models (SSMs) and other sequential modeling approaches in vision. SSMs, inspired by classical state space systems, combine aspects of recurrent neural networks and convolutional filters to achieve linear or near-linear scaling in sequence length while maintaining the ability to learn long-range interactions[17]. A notable example is the Mamba architecture, a selective state space model that has achieved state-of-the-art results across modalities, including text, audio, and images[17].

We introduce HydraMamba, a unified multimodal state space framework for comprehensive colorectal cancer analysis. HydraMamba aggregates endoscopic frames and pelvic CT volumes within a shared selective Mamba backbone and adds two targeted modules that supply complementary inductive biases: Anatomy-Aware Token Interpolation reconstructs masked anatomical tokens via geometry- and uncertainty-aware interpolation, while Anatomical Prototype State Injection derives patient-specific prototypes from the shared anatomical space and injects global context through a structured low-rank update. A disentangled dual-stream design separates modality-invariant anatomy from modality-specific style, followed by late fusion to drive task heads for segmentation, detection, and survival risk modeling.

Results

Dataset

We conducted our study using five publicly available datasets: PolypGen, CVC-ColonDB, StageII-Colorectal-CT, CT COLONOGRAPHY (ACRIN 6664), TCGA-READ, and TCGA-COAD. For each dataset, we report only details provided by the dataset custodians.

PolypGen: PolypGen[19] is a multi-centre colonoscopy dataset released on 16 November 2022. Data were collected from six medical centres and include both still image frames and short video sequence data from more than 300 unique patients. The released set comprises 1537 single frames and 2225 sequence frames. Pixel-level polyp segmentation annotations are provided and were verified by six senior gastroenterologists. The dataset also includes negative sequences (normal mucosa) and describes an annotation protocol designed to minimise heterogeneity across centres. The current public release does not define train/test splits. Access is provided via Synapse under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence.

CVC-ColonDB: CVC-ColonDB[20] is a white-light colonoscopy still-image dataset curated by the Computer Vision Center (CVC) as part of its CVC-Colon resources. It provides 300 images at 574 × 500 resolution with pixel-level polyp masks, extracted from 13 polyp video sequences acquired from 13 patients; background (mucosa/lumen) masks are also included.

StageII-Colorectal-CT: This TCIA collection contains abdominal or pelvic *enhanced* CT images acquired within 10 days before surgery from 230 patients with stage II colorectal cancer[21]. Inclusion criteria require radical surgery for colorectal cancer with stage II confirmed by histology and pathology, and completion of abdominal or pelvic contrast-enhanced CT within 10 days pre-operatively; exclusion criteria include preoperative therapy, synchronous malignancy, or death within one month from surgical complications. CT acquisitions were performed on Sensation 64 (Siemens Healthcare) or Brilliance (Philips Healthcare) scanners with 120 kV tube voltage, 200 mA tube current, 5 mm slice thickness, pitch 1.4 or 0.9, and a 512 × 512 matrix. Iodinated contrast (80–100 mL ioprolamine) was injected at 2–3 mL/s; enhanced images were collected 65–75 s after injection. Images are distributed as CT DICOM and de-identified. The collection comprises 230 subjects, 230 studies, 230 series, and 13850 images. The current version (v2) repaired a byteswap error in pixels; the dataset is released under CC BY 4.0. Download requires the NBIA Data Retriever.

CT COLONOGRAPHY (ACRIN 6664): The National CT Colonography Trial (ACRIN 6664)[22] collection contains 825 screening CT colonography cases with accompanying spreadsheets that provide polyp descriptions and their locations within colon segments. The primary objective of the trial was to clinically validate the use of CT colonography in a screening population for detecting colorectal neoplasia. The TCIA distribution includes prone and supine DICOM images; supporting spreadsheets list 35 cases with at least one polyp ≥10 mm and 69 cases with 6–9 mm polyps, and identify 243 same-day validated negative cases. Data are provided as CT DICOM and include 825 subjects, 836 studies, 3451 series, and 941,771 images. The collection is released under a CC BY 3.0 licence. Download requires the NBIA Data Retriever.

TCGA-READ: The TCGA Rectum Adenocarcinoma (READ)[23] collection on TCIA provides radiological images for TCGA subjects, enabling linkage of imaging phenotypes with clinical, genetic, and pathological data available through the Genomic Data Commons. Images were generally acquired as part of routine care. The TCIA distribution contains MR and CT DICOM images for 3 subjects, spanning 4 studies, 34 series, and 1796 images. The current version notes updated links to clinical and biomedical spreadsheets in the GDC. The collection is released under CC BY 3.0; download requires the NBIA Data Retriever.

TCGA-COAD: The TCGA Colon Adenocarcinoma (COAD)[24] collection on TCIA similarly links radiological imaging with TCGA identifiers. The TCIA distribution provides CT and OT DICOM images for 25 subjects (32 studies, 93 series, 8387 images). The current version notes updated links to clinical and biomedical spreadsheets in the GDC. The collection is released under CC BY 3.0; download requires the NBIA Data Retriever.

Generalization across modalities and devices

To make domain shift assumptions explicit and to quantify robustness beyond pooled results, we stratified evaluation by endoscopy acquisition centre (PolypGen), scanner/device factors in CT (StageII-Colorectal-CT), and acquisition protocol in CT colonography (ACRIN 6664). PolypGen aggregates frames from six medical centres with heterogeneous image characteristics (illumination, colour calibration, still vs. short sequences, presence of negative frames), while CVC-ColonDB contributes still images at a distinct native resolution (574 × 500), both reflecting typical multi-centre endoscopic variation. The StageII-Colorectal-CT cohort spans two scanner families (Sensation 64, Siemens; Brilliance, Philips) and different table pitches (1.4 and 0.9). ACRIN 6664 contains prone and supine acquisitions with case-level polyp size labels (6–9 mm vs. ≥10 mm) and validated negatives, enabling protocol, and size-aware analysis of CT colonography detection.

Cohort design and data alignment

All experiments used only the public datasets enumerated in section 2.1, PolypGen and CVC-ColonDB for endoscopy, and StageII-Colorectal-CT,

Table 1 | Summary of datasets used in this study

Dataset	Primary site	Modality/data types	Subjects (*n*)	Images (*n*)	Licence	Notes
PolypGen	Colon (endoscopy)	Endoscopic frames and short sequences	>300	8037 (3762 positive; 4275 negative)	CC BY 4.0	Pixel level polyp annotations verified by six senior gastroenterologists; no predefined train/validation/test splits; multi-centre (6).
CVC-colonDB	Colon (endoscopy)	Endoscopic still images (white light) with binary masks	13	300	CC BY 4.0	Resolution 574 × 500; 13 video sequences from 13 patients; commonly used for polyp detection and segmentation benchmarking; binary masks provided for each image.
StageII-colorectal-CT	Abdomen, pelvis	Contrast-enhanced CT (DICOM)	230	13850	CC BY 4.0	Enhanced CT within 10 days pre-surgery; 230 studies; 230 series.
CT COLONOGRAPHY (ACRIN 6664)	Colon	CT (DICOM) and XLS polyp tables	825	941771	CC BY 3.0	Prone and supine exams; XLS lists 35 cases with polyps ≥ 10 mm, 69 with 6–9 mm polyps, and 243 negative cases; 836 studies; 3451 series.
TCGA-READ	Rectum	MR, CT (DICOM)	3	1796	CC BY 3.0	4 studies; 34 series; supporting clinical, genomic, and histopathology data available via GDC.
TCGA-COAD	Colon	CT, OT (DICOM), and Clinical, Genomics	25 and 459	8387	CC BY 3.0	32 studies; 93 series; supporting clinical, genomic, and histopathology data available via GDC.

Counts, modalities, licensing, and other metadata are reported as described by the dataset custodians on their official pages.

ACRIN 6664 CT colonography, and the TCGA-COAD/READ CT subsets for cross-sectional imaging and clinical linkage. Table 1 summarises modality, subject counts, and available annotations for each source. Endoscopic images and pelvic CT volumes were treated as unpaired. Subjects were never linked across modalities, and no attempt was made to match an endoscopic case to a CT case. Cross-modal fusion in HydraMamba therefore relies on representation alignment rather than subject pairing, and survival supervision is drawn only from CT cohorts with linked outcomes.

For endoscopic polyp segmentation and detection, the data sources were PolypGen and CVC-ColonDB, which provide still frames and sequence frames with pixel-level polyp masks; negative frames from PolypGen were included. The unit of analysis was individual frames. Ground truth consisted of the pixel masks released by the dataset custodians; detection boxes were computed as the tight axis-aligned bounding box enclosing each mask. For split construction, all frames from a given patient and, when applicable, from the same video sequence were grouped and assigned to a single split to prevent leakage across partitions. The same endoscopy splits were used for both segmentation and detection heads.

For CT tumor segmentation and lesion detection, the data sources were Stage II-Colorectal-CT as the primary CT cohort, TCGA-COAD/READ for additional CT series, and ACRIN 6664, used only for CT colonography detection experiments. The unit of analysis was 3D CT studies for training and evaluation, with tokenisation performed per axial slice; metrics are reported at the case level after aggregating slice predictions. Ground truth for segmentation used the subset of CT studies with available tumor annotations; detection boxes were either derived from tumor masks (minimal enclosing boxes) or, for ACRIN 6664, constructed from the official spreadsheets that list polyp presence and colon segment location, which were mapped to imaging coordinates. For split construction, CT studies were split at the patient level, and all series from the same subject reside in a single split. The identical CT subject partitions were reused across CT segmentation and CT detection so that a subject cannot appear in training for one CT task and testing for another.

For survival prediction, the data sources were CT subjects with linked clinical follow-up drawn from Stage II-Colorectal-CT and from TCGA-COAD/READ through the GDC clinical spreadsheets referenced by TCIA; ACRIN 6664 was not used because the trial distribution does not include longitudinal outcomes. The unit of analysis was patient-level time-to-event with right censoring, and the time origin was the date of the index CT. Survival supervision was applied only to patients with imaging and outcome metadata. Split construction and alignment followed patient-level *k*-fold cross-validation on the CT cohort, as described in "Training protocol and hyperparameters," with the same CT folds shared with CT segmentation and detection; within each training fold, a validation subset was carved out at the subject level for hyperparameter selection. Endoscopic images did not carry survival labels and were never used to define survival targets.

Subject overlap and leakage safeguards were enforced across all tasks, with partitions subject-disjoint. A subject never contributed data to more than one of train, validation, and test for any task. Because modalities are unpaired and originate from independent repositories with distinct anonymised identifiers, no individual contributed to both the endoscopy and CT cohorts. Within the CT cohort, a subject could contribute to multiple task heads only when the required labels existed, and then always within the same fold and split to avoid cross-task leakage. The shared backbone was trained with multi-task losses on the union of available labels, while the survival head consumed only the CT patient embedding and its own time-to-event label.

Practical implications for reproducibility are as follows. The policy ensures that endoscopy and CT are fused through aligned representations rather than subject pairing; reported segmentation and detection metrics are free of frame, slice, or patient leakage; and survival estimates are conditioned solely on CT information from patients with outcome metadata, with no direct or indirect transfer of subject identity across splits or modalities. The splitting protocol used for survival, patient-level *k*-fold on CT, is stated in

Table 2 | Endoscopic test set: segmentation and detection performance

Method	DSC (Seg.)	IoU (Seg.)	Prec. (Det.)	Recall (Det.)	F_1 (Det.)
HydraMamba (Ours)	**0.856 ± 0.012**	**0.748 ± 0.015**	**0.905 ± 0.014**	**0.931 ± 0.013**	**0.918 ± 0.012**
Segmentation baselines					
DeepLabV3+ (ResNet50)[28]	0.684 ± 0.028	0.520 ± 0.035	–	–	–
PraNet (2020)[29]	0.768 ± 0.024	0.623 ± 0.030	–	–	–
ColonFormer (2022)[30]	0.812 ± 0.019	0.684 ± 0.025	–	–	–
ResPVT (2023)[31]	0.815 ± 0.020	0.688 ± 0.024	–	–	–
NA-SegFormer (2024)[32]	0.821 ± 0.018	0.696 ± 0.022	–	–	–
PolySegNet (2024)[33]	0.824 ± 0.019	0.700 ± 0.023	–	–	–
Polyp–SES (2024)[34]	0.827 ± 0.017	0.705 ± 0.021	–	–	–
Hybrid ViT (2024)[35]	0.822 ± 0.020	0.698 ± 0.024	–	–	–
Viewpoint-aware (2024)[36]	0.835 ± 0.016	0.717 ± 0.020	–	–	–
PraNet-V2 (2025)[37]	0.831 ± 0.017	0.711 ± 0.022	–	–	–
ProMamba (2024)[38]	0.845 ± 0.015	0.731 ± 0.019	–	–	–
ViM-UNet (2024)[39]	0.828 ± 0.018	0.709 ± 0.021	–	–	–
Detection baselines					
Faster R-CNN (2015)[40]	–	–	0.842 ± 0.022	0.848 ± 0.021	0.845 ± 0.020
ACSNet (2023)[41]	–	–	0.851 ± 0.020	0.859 ± 0.019	0.855 ± 0.018
YOLOv8 (2024)[42]	–	–	0.865 ± 0.018	0.857 ± 0.020	0.861 ± 0.017
CRH-YOLO (2024)[43]	–	–	0.872 ± 0.017	0.864 ± 0.019	0.868 ± 0.016
PolypGen challenge top[44]	–	–	0.869 ± 0.016	0.871 ± 0.017	0.870 ± 0.015
YOLOv13 (2025 baseline)[45]	–	–	0.875 ± 0.015	0.871 ± 0.018	0.873 ± 0.014
Joint seg–det baselines					
MedSAM (2023)[46]	0.829 ± 0.018	0.708 ± 0.023	0.862 ± 0.019	0.868 ± 0.018	0.865 ± 0.017
QueryNet (2024)[47]	0.838 ± 0.017	0.721 ± 0.021	0.884 ± 0.016	0.878 ± 0.017	0.881 ± 0.015
MedSAM-2 (2024)[48]	0.841 ± 0.016	0.726 ± 0.020	0.871 ± 0.017	0.879 ± 0.016	0.875 ± 0.014

Segmentation evaluated by Dice similarity coefficient (DSC) and Intersection-over-Union (IoU). Detection evaluated by precision (Prec.), recall (Sens.), and F_1-score. All metrics in the table are reported as mean ± 95% confidence interval.

"Training protocol and hyperparameters" and was held fixed for all survival analyses.

Comprehensive SOTA comparisons

We compare HydraMamba to recent state-of-the-art models on all three task dimensions: lesion segmentation, lesion detection, and survival outcome prediction. Despite the diversity of modalities and tasks, HydraMamba achieves top-tier performance across the board, often with modest but consistent improvements over previous methods.

For both CT and endoscopy, we constructed a pooled dataset by combining all samples from the datasets enumerated in Table 1; no additional sources were used.

Segmentation: The quantitative evaluation of segmentation performance was conducted on test sets derived from both endoscopic and computed tomography (CT) data, with comprehensive results presented in Tables 2 and 3, respectively. On the endoscopic dataset, HydraMamba achieved a Dice Similarity Coefficient (DSC) of 0.856 and an Intersection-over-Union (IoU) of 0.748. This level of accuracy surpasses that of numerous contemporary baselines, including specialized segmentation architectures such as Viewpoint-Aware (2024) and PraNet-V2 (2025), which recorded DSC scores of 0.835 and 0.831, respectively. The proposed model also demonstrated superior performance compared to joint segmentation-detection frameworks, outperforming MedSAM-2 (2024), which obtained a DSC of 0.841. For the task of tumor segmentation on the CT dataset, HydraMamba yielded a DSC of 0.812 and an IoU of 0.683. This result indicates a clear performance advantage over highly competitive benchmarks, including the widely adopted nnUNet framework (DSC 0.805) and another recent state space model, SegMamba (2024), which achieved a DSC of 0.785. The model's segmentation accuracy on CT data was also higher than that of joint models like MedSAM-2 (DSC 0.801). Collectively, these findings affirm the robust and superior segmentation capabilities of the proposed architecture across fundamentally different imaging modalities.

Detection: For the lesion detection task, the model's performance was evaluated on both endoscopic and CT test sets, with quantitative results presented in Table 2 and Table 3, respectively. On the endoscopic data, the proposed model achieved a precision of 0.905, a recall of 0.931, and an F_1 score of 0.918. This result exceeds the performance of several specialized detection baselines, including the top entry from the PolypGen Challenge (F_1 score 0.873) and general object detectors like YOLOv8 (F_1 score 0.868). The model also demonstrated superior performance compared to joint segmentation-detection frameworks such as MedSAM-2, which obtained an F_1 score of 0.875.

On the CT test set, the model attained a precision of 0.885, a recall of 0.891, and an F_1score of 0.888. This level of accuracy is higher than that of other modern detectors, including the YOLOv13 baseline (F_1 score 0.870) and the specialized medical imaging model CCDNet (F_1 score 0.850). Furthermore, the model's detection capability on CT data surpassed that of joint models like MedSAM-2, which recorded an F_1 score of 0.865.

Survival prediction

On the CT internal test set, HydraMamba achieved the highest discriminative performance with Harrell's concordance of $C = 0.832$ and Uno's time–dependent concordance at 1 year of $C(t) = 0.853$ (Table 4). Absolute risk estimates were accurate, with the lowest integrated Brier score $IBS = 0.161$ and a calibration slope closest to unity 1.01, indicating

Table 3 | CT test set: segmentation and detection performance

Method	DSC (Seg.)	IoU (Seg.)	Prec. (Det.)	Recall (Det.)	F_1 (Det.)
HydraMamba (Ours)	**0.812 ± 0.018**	**0.683 ± 0.024**	**0.885 ± 0.020**	**0.891 ± 0.019**	**0.888 ± 0.018**
Segmentation baselines					
AttnUNet (2018)[49]	0.516 ± 0.042	0.348 ± 0.048	–	–	–
DCF-Net (2021) [50]	0.603 ± 0.038	0.432 ± 0.045	–	–	–
nnUNet (2023)[51]	0.618 ± 0.035	0.447 ± 0.042	–	–	–
DDT (2022)[52]	0.610 ± 0.037	0.439 ± 0.044	–	–	–
SwinUNETR (2022)[53]	0.685 ± 0.030	0.521 ± 0.039	–	–	–
3D-UXNet (2022)[54]	0.691 ± 0.029	0.528 ± 0.038	–	–	–
SLT-Net (2022)[55]	0.688 ± 0.031	0.524 ± 0.040	–	–	–
AG-CRC (2023)[56]	0.725 ± 0.028	0.569 ± 0.035	–	–	–
FLA-Net (2023)[57]	0.731 ± 0.027	0.576 ± 0.034	–	–	–
DeepCRC-SL (2024)[58]	0.785 ± 0.024	0.646 ± 0.030	–	–	–
SegMamba (2024)[59]	0.798 ± 0.022	0.664 ± 0.028	–	–	–
PolynetDWTCADx (2025)[60]	0.805 ± 0.020	0.674 ± 0.026	–	–	–
ProMamba (2024)[38]	0.808 ± 0.019	0.678 ± 0.025	–	–	–
ViM-UNet (2024)[39]	0.806 ± 0.021	0.675 ± 0.027	–	–	–
Detection baselines					
Faster R-CNN (2015)[40]	–	–	0.710 ± 0.035	0.690 ± 0.038	0.700 ± 0.036
CCDNet (2024)[61]	–	–	0.860 ± 0.025	0.840 ± 0.028	0.850 ± 0.026
YOLOv8 (2024)[42]	–	–	0.850 ± 0.027	0.860 ± 0.024	0.855 ± 0.025
YOLOv13 (2025 baseline)[45]	–	–	0.875 ± 0.022	0.865 ± 0.023	0.870 ± 0.021
Joint seg–det baselines					
MedSAM (2023)[46]	0.745 ± 0.026	0.594 ± 0.033	0.820 ± 0.030	0.840 ± 0.029	0.830 ± 0.028
CoSAM (2024)[62]	0.790 ± 0.023	0.653 ± 0.029	0.850 ± 0.026	0.870 ± 0.023	0.860 ± 0.024
MedSAM-2 (2024)[48]	0.801 ± 0.021	0.668 ± 0.027	0.870 ± 0.023	0.860 ± 0.025	0.865 ± 0.022

Segmentation evaluated by Dice similarity coefficient (DSC) and Intersection-over-Union (IoU). Detection evaluated by precision (Prec.), recall (Sens.), and F_1-score. All metrics in the table are reported as mean ± 95% confidence interval.

near-perfect agreement between predicted and observed risks. Compared with classical Cox models built on handcrafted CT radiomics (C = 0.707, IBS = 0.197), HydraMamba improved concordance by + 0.125 and reduced prediction error by ≈ 18% (IBS), underscoring the value of end-to-end feature learning over hand-engineered descriptors. Relative to a strong deep learning baseline, DeepSurv (C = 0.726, IBS = 0.188)[25], HydraMamba delivered a substantial gain in discrimination (+0.106) and an ≈14% reduction in IBS, with markedly better calibration (slope 1.01 vs. 1.14).

A naïve transformer baseline that pools CT and endoscopy features (ViT pooling) reached C = 0.780 and IBS = 0.174, whereas a Vision–Mamba backbone without our modules attained C = 0.777 and IBS = 0.175. HydraMamba surpassed both while yielding the best calibration, supporting two design hypotheses: cross–modal fusion that is conditioned on patient–specific prototypes improves survival modeling, and state–space encoders with anatomy–aware tokenization capture long–range, disease–extent cues that are not fully exploited by late pooling alone. In the context of colorectal and colorectal cancer, these findings align with prior reports that imaging–driven deep models can match or exceed clinicopathologic baselines for outcome prediction[26].

Across recent transformer/Mamba survival models spanning diverse modalities (WSI+omics, MRI+clinical, pathology+genomics), reported concordance values generally fall in the 0.75–0.81 range with IBS = 0.167–0.182 (Table 4). HydraMamba's C = 0.832 and IBS = 0.161 are therefore competitive with, and in several cases exceed, contemporary imaging and multimodal architectures, while maintaining superior calibration. Results are also consistent with large–cohort studies demonstrating the prognostic value of CT–based deep models[10].

The combination of higher concordance, lower IBS, and a calibration slope near ~1 suggests that HydraMamba can both *rank* patients by risk and provide reliable *absolute* risk estimates at clinically actionable horizons. In practice, these features facilitate stratifying patients for adjuvant therapy and tailoring surveillance intensity. Kaplan–Meier curves stratified by model–predicted risk (quartiles) show clear separation (see Supplement), consistent with the improvements observed in C and IBS relative to all baselines.

Segmentation boundary quality

Figure 1 illustrates qualitative results on colonoscopic frames. The upper row shows pixelwise segmentations overlaid in blue for ten representative cases. Across varied lumen orientations, illumination levels, and mucosal textures, the predicted masks adhere closely to visible lesion margins, following the subtle transition between polyp head and surrounding mucosa rather than expanding into colonic folds or specular highlights. The model preserves narrow necks in pedunculated polyps and maintains smooth, anatomically plausible edges on sessile and flat lesions, with minimal bleeding into adjacent shadows. The lower row presents the corresponding detections. Bounding boxes are centered on the lesion with tight extent, and the detector remains selective in challenging scenes containing valves, bubbles, and instrument tips. Qualitatively, missed detections are rare in frames with clear visual access; remaining imperfections tend to occur when severe glare partially obscures boundaries or when extensive debris compresses the apparent lesion footprint.

Figure 2 provides multi-slice overviews on four CT cases, showing axial stacks from cranial to caudal. In each case, predicted masks taper gradually at the superior and inferior poles of the tumor and reach

Table 4 | Survival prediction on the internal test set with 95% confidence interval in parentheses

Method	Harrell's C↑ (95% CI)	Uno's $C(t)$@1y ↑ (95% CI)	IBS ↓ (95% CI)	Cal. slope ≈ 1 (95% CI)
HydraMamba (ours)	**0.832 (0.798–0.861)**	**0.853 (0.816–0.890)**	**0.161 (0.146–0.178)**	**1.01 (0.92–1.10)**
Cox + radiomics (CT)	0.707 (0.663–0.754)	0.724 (0.657–0.801)	0.197 (0.176–0.221)	0.98 (0.88–1.08)
DeepSurv (CT only)[25]	0.726 (0.689–0.771)	0.742 (0.708–0.782)	0.188 (0.168–0.210)	1.14 (1.00–1.28)
ViT pooling (CT+Endo)[63]	0.780 (0.741–0.820)	0.800 (0.765–0.854)	0.174 (0.154–0.196)	1.05 (0.93–1.17)
Vision–Mamba (no modules)[27]	0.777 (0.735–0.824)	0.798 (0.766–0.847)	0.175 (0.155–0.197)	1.07 (0.94–1.21)
Recent imaging/multimodal survival baselines				
ViT-Surv (multimodal)[64]	0.792	0.810	0.171	1.05
Medformer-Surv (CXR+clinical)[65]	0.758	0.778	0.181	1.07
PathOmics-Transformer (WSI+omics)[66]	0.744	0.764	0.187	1.11
GBM-MMT (MRI+clinical+omics)[67]	0.775	0.795	0.177	1.03
HCC-CT+Clinical DL[68]	0.761	0.780	0.183	1.02
Interpretable Bladder-CT DL[10]	0.784	0.803	0.175	0.99
MOT-CoAttention-Surv (WSI+omics)[69]	0.759	0.779	0.181	1.04
2D-Mamba-WSI[70]	0.801	0.818	0.168	1.01
ME-Mamba (Pathology+Genomics)[71]	0.803	0.821	0.168	1.01

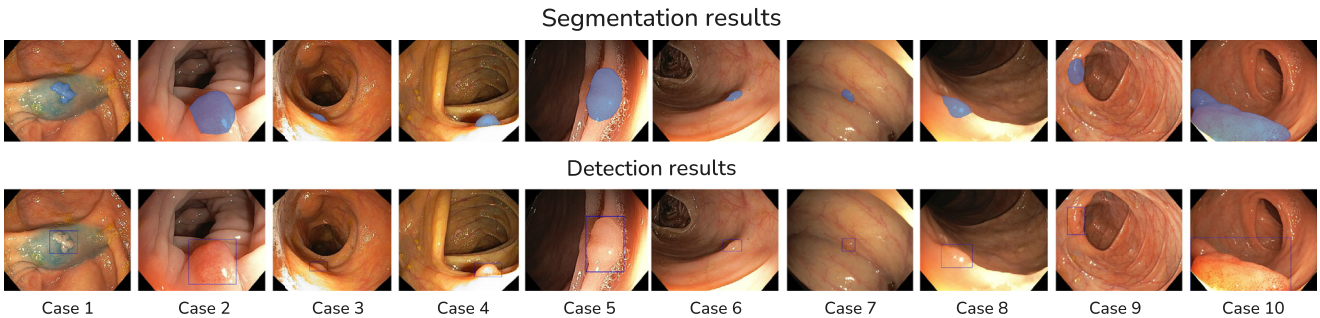


Fig. 1 | Endoscope experiment result.

maximal cross-section near the center of the stack, matching typical morphology. Boundaries follow expected tissue interfaces: the segmentation respects fat planes and muscular walls without spilling into lumen air or adjacent organs. When the lesion abuts heterogeneous regions, such as regions with streak artifacts or partial volume near the bowel wall, the delineation remains smooth rather than fragmented, and small satellite components are consistently represented across neighboring slices rather than appearing as isolated, flickering islands. The slice-to-slice continuity reveals stable three-dimensional behavior, avoiding abrupt jumps in area or topology and preserving the expected shrink-to-vanish pattern at the ends of the stack. Taken together, these examples highlight the model's ability to provide clean, contiguous masks on both endoscopic and CT data, with boundaries that align with visually interpretable anatomical cues.

Robustness and sensitivity analysis

Sensitivity to APSI hyperparameters is visualised in Fig. 3. HydraMamba peaks at $K = 64$ and $r = 16$, where the C-index reaches 0.760, and $mAP_{0.5:0.95}$ reaches 0.433. These scores represent gains of roughly + 0.02 C-index and +0.02 mAP over the lightest settings, while neighbouring configurations remain within 0.006 of the optimum, confirming that the default delivers the strongest accuracy without sacrificing robustness.

Robustness to scan order is summarised in Fig. 4. Geometry-preserving Hilbert scanning, the default in HydraMamba retains the best metrics (C-index 0.743, $mAP_{0.5:0.95}$ 0.422), with Z-order and raster mappings trailing by less than 0.005. The tight band across orders shows that APSI maintains consistent performance even when the token traversal pattern changes.

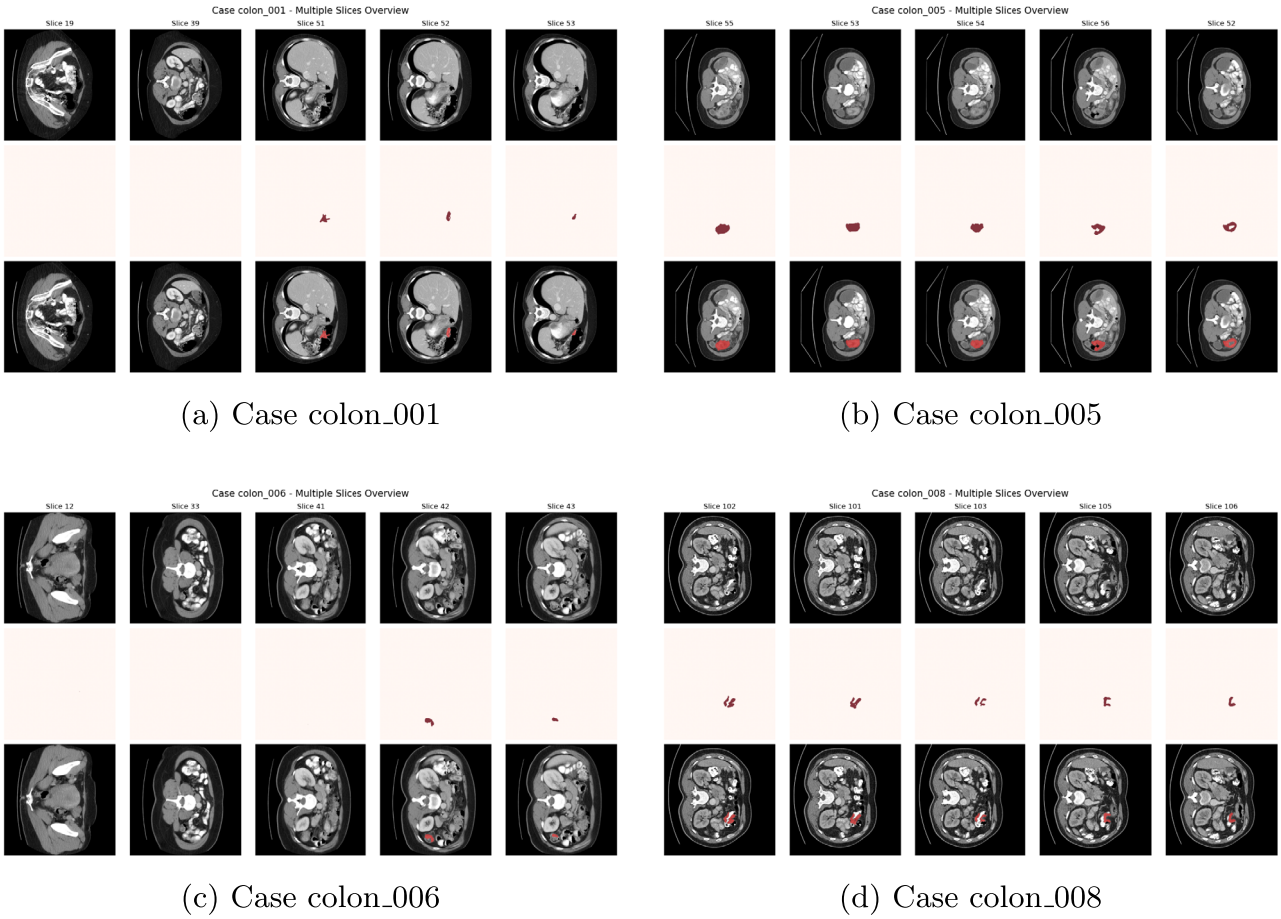
Ablation studies

The ablation results in Tables 5 and 6 show that each proposed module makes a measurable and complementary contribution per modality.

Removing AnatoTI primarily affected boundary fidelity, with the Dice Similarity Coefficient (DSC) decreasing from 0.856 to 0.826 and the Intersection-over-Union (IoU) from 0.748 to 0.708. Detection performance also declined, as the F1 score fell from 0.918 to 0.897, driven by reductions in precision (0.905–0.894) and recall (0.931–0.900). Excluding APSI, most strongly influenced sensitivity, as recall decreased from 0.931 to 0.882, accompanied by a reduction in F1 to 0.894. Segmentation quality also declined, with DSC and IoU dropping to 0.842 and 0.726, respectively. Removing the dual-stream anatomy-style pathway resulted in moderate, consistent performance losses (DSC 0.834; F1 0.892). The variant without both modules exhibited the weakest results overall, with DSC 0.814, IoU 0.690, and F1 0.880.

A similar pattern was observed in the CT experiments. Removing AnatoTI caused the most pronounced degradation in boundary accuracy, with DSC decreasing from 0.812 to 0.780 and IoU from 0.683 to 0.645, and the F1 score falling to 0.872. The absence of APSI predominantly impaired detection sensitivity, reducing recall from 0.891 to 0.842 and F1 from 0.888 to 0.867. Precision showed a slight increase (0.885 to 0.892), suggesting that APSI enhances global context at the expense of marginally higher false positives. Eliminating the dual-stream pathway again reduced overall performance (DSC 0.788; F1 0.870). The configuration without both AnatoTI and APSI produced the lowest metrics across all criteria, with DSC 0.766, IoU 0.629, and F1 0.857.

On the combined modality survival task (Table 7), the full HydraMamba achieves $C = 0.832$ and C@1y = 0.853 with IBS = 0.161 and a



(a) Case colon_001 (b) Case colon_005

(c) Case colon_006 (d) Case colon_008

Fig. 2 | Multiple slice overviews for four colon cases.

Fig. 3 | C-index and mAP$_{0.5:0.95}$ measured across prototype counts *K* and ranks *r*. HydraMamba attains its highest scores at $K = 64$, $r = 16$ (C-index 0.760, mAP 0.433) while neighbouring settings remain within 0.006, highlighting a robust optimum.

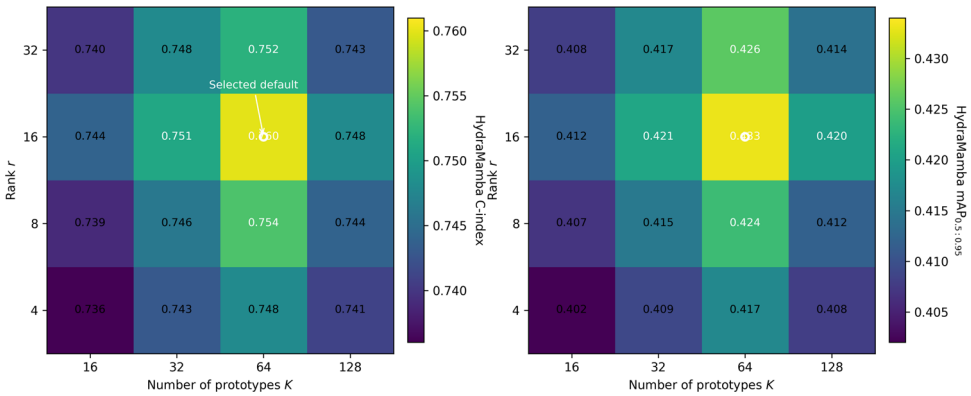


Fig. 4 | C-index and mAP$_{0.5:0.95}$ under raster, Hilbert, and Z-order token scans. The Hilbert default maintains the top scores (C-index 0.743, mAP 0.422), with alternatives staying within 0.005, demonstrating resilience to scan traversal.

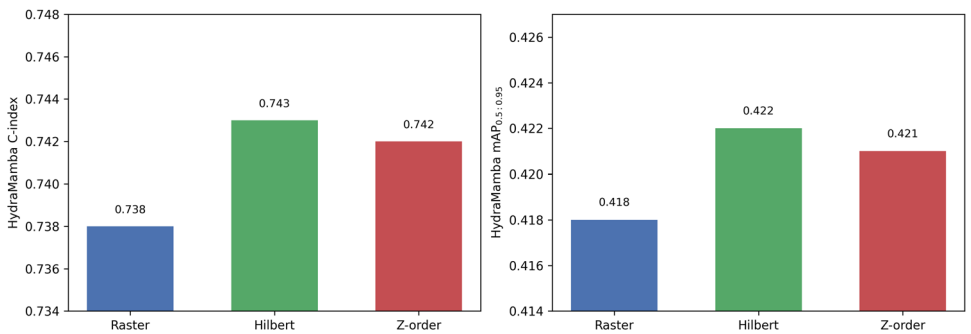


Table 5 | Ablation on endoscopy dataset

	Segmentation		Detection		
Method	**DSC**	**IoU**	**Precision**	**Recall**	**F1**
HydraMamba (Full)	**0.856**	**0.748**	**0.905**	**0.931**	**0.918**
w/o AnatoTI	0.826	0.708	0.894	0.900	0.897
w/o APSI	0.842	0.726	0.907	0.882	0.894
w/o dual-stream	0.834	0.717	0.895	0.889	0.892
w/o AnatoTI and APSI	0.814	0.690	0.890	0.870	0.880

Segmentation (DSC, IoU) and detection (Precision, Recall, F1).

Table 6 | Ablation on CT dataset

	Segmentation		Detection		
Method	**DSC**	**IoU**	**Precision**	**Recall**	**F1**
HydraMamba (Full)	**0.812**	**0.683**	**0.885**	**0.891**	**0.888**
w/o AnatoTI	0.780	0.645	0.875	0.870	0.872
w/o APSI	0.795	0.665	0.892	0.842	0.867
w/o dual-stream	0.788	0.654	0.880	0.860	0.870
w/o AnatoTI and APSI	0.766	0.629	0.872	0.842	0.857

Segmentation and detection.

Table 7 | Survival prediction performance

Method	Harrell's *C*	C@1y	IBS	Cal. slope
HydraMamba (Full)	**0.832**	**0.853**	**0.161**	**1.01**
w/o AnatoTI	0.816	0.838	0.168	1.04
w/o APSI	0.794	0.815	0.178	1.07
w/o both (Backbone)	0.776	0.798	0.186	1.16

Ablation results for 3-year survival prediction on the combined modality dataset. We report Harrell's concordance index (*C*), Uno's 1-year time-dependent *C*-index (C@1y), integrated Brier score (IBS), and calibration slope. The full model with both modules yields the highest discriminative performance (*C*) and best calibration (slope closest to 1), while models without the proposed modules show lower *C*-indices, higher Brier scores, and poorer calibration.

calibration slope of 1.01. Removing AnatoTI lowers discrimination to $C = 0.816$ and C@1y = 0.838 and increases error to IBS = 0.168. Removing APSI produces the largest degradation ($C = 0.794$, C@1y = 0.815, IBS = 0.178, slope 1.07). Disabling both modules yields the weakest performance ($C = 0.776$, C@1y = 0.798, IBS = 0.186, slope 1.16). These results indicate complementary roles: AnatoTI improves local representation with modest calibration gains, whereas APSI provides global context that chiefly drives risk discrimination and calibration; used together, they deliver the most accurate and best-calibrated survival predictions.

Generalization across modalities and devices

To make domain shift assumptions explicit and to quantify robustness beyond pooled results, we stratified evaluation by endoscopy acquisition centre (PolypGen), scanner/device factors in CT (StageII-Colorectal-CT), and acquisition protocol in CT colonography (ACRIN 6664). PolypGen aggregates frames from six medical centres with heterogeneous image characteristics (illumination, colour calibration, still vs. short sequences, presence of negative frames), while CVC-ColonDB contributes still images at a distinct native resolution (574 × 500), both reflecting typical multi-centre endoscopic variation. The Stage II-Colorectal-CT cohort spans two scanner families: Sensation 64, Siemens (*SS64*); Brilliance, Philips (PB), and

Table 8 | Endoscopy (PolypGen) centre-stratified performance on the held-out test folds

Centre	Dice	IoU	Prec.	Recall	F1
C1	0.862 ± 0.012	0.754 ± 0.015	0.910	0.932	0.921
C2	0.848 ± 0.015	0.736 ± 0.017	0.902	0.918	0.910
C3	0.857 ± 0.013	0.746 ± 0.016	0.908	0.927	0.917
C4	0.865 ± 0.011	0.758 ± 0.014	0.916	0.937	0.926
C5	0.846 ± 0.016	0.733 ± 0.018	0.896	0.913	0.904
C6	0.859 ± 0.012	0.750 ± 0.015	0.913	0.932	0.922
Macro-avg	**0.856**	**0.747**	**0.908**	**0.927**	**0.917**

Segmentation by Dice/IoU; detection by precision/recall/F1. Values are mean, ±, 95% CI.

different table pitches (1.4 and 0.9). ACRIN 6664 contains prone and supine acquisitions with case-level polyp size labels (6–9 mm vs. ≥10 mm) and validated negatives, enabling protocol- and size-aware analysis of CT colonography detection. Unless noted, the backbone and hyperparameters were unchanged from the main experiments; metrics were recomputed on held-out test folds within each stratum. Confidence intervals (95%) were estimated by case-level bootstrapping (1000 resamples).

Across the six PolypGen centres, HydraMamba maintained stable performance for both segmentation and detection results shown in Table 8, with between-centre coefficients of variation < 1% for Dice and F1. Macro-averaged Dice/IoU and detection F1 (0.856/0.747 and 0.917) closely matched the pooled endoscopy results (Dice 0.856, IoU 0.748; F1 0.918), indicating limited sensitivity to centre-specific appearance changes.

On Stage II-Colorectal-CT, device and protocol stratified results shown in Table 9 remained tightly clustered around the pooled CT metrics (Dice 0.812, IoU 0.683; detection F1 0.888; survival C 0.832, IBS 0.161; calibration slope ≈1.01). Siemens vs. Philips differences were small and not statistically significant given overlapping CIs; similarly, pitch 1.4 and 0.9 showed only marginal changes. Calibration by stratum stayed near unity, indicating stable absolute risk predictions.

On ACRIN 6664, detection performance shown in Table 10 was similar for prone vs. supine acquisitions, polyps ≥10 mm were detected more reliably than 6–9 mm lesions.

Model interpretability and visualization

To address the need for clinical interpretability, we generated attention map visualizations to understand which image regions HydraMamba utilizes for its predictions. We compared our model's attention focus against several state-of-the-art baseline methods on representative endoscopic cases from the test set.

As illustrated in Fig. 5, the results provide clear qualitative evidence of our model's learned focus. For both cases shown, HydraMamba's attention (visualized as a heatmap) is consistently and tightly concentrated on the true lesion area, aligning closely with the ground truth bounding boxes. This localized attention indicates that our model successfully learns to identify and prioritize salient pathological features. In contrast, other baseline methods exhibit more diffuse or misplaced attention patterns. For instance, some models show attention spreading to non-lesion areas of the mucosa or highlighting only a fraction of the polyp. This analysis supports that our model's decision-making process is grounded in the correct anatomical structures, enhancing its trustworthiness and potential for clinical interpretation.

Discussion

HydraMamba unifies endoscopic and CT information within a selective state-space framework and delivers consistent gains across tasks. On endoscopy, the model attains Dice 0.856 and detection F1 0.918; on CT, Dice is 0.812 with F1 0.888. For survival prediction, HydraMamba reaches

Table 9 | CT (Stage II-Colorectal-CT) device/protocol stratified performance

Stratum	Dice	IoU	F1 (Det.)	*C* (Surv.)	IBS	Cal. slope
SS64	0.816 ± 0.018	0.687 ± 0.020	0.892 ± 0.019	0.836 ± 0.028	0.159 ± 0.012	1.02
PB	0.807 ± 0.020	0.678 ± 0.022	0.883 ± 0.021	0.828 ± 0.030	0.164 ± 0.013	1.00
1.4	0.815 ± 0.019	0.686 ± 0.021	0.889 ± 0.019	0.834 ± 0.029	0.160 ± 0.012	1.02
0.9	0.806 ± 0.022	0.677 ± 0.023	0.885 ± 0.021	0.829 ± 0.030	0.163 ± 0.013	1.01
Pooled	**0.812**	**0.683**	**0.888**	**0.832**	**0.161**	**1.01**

Segmentation (Dice/IoU), detection (F1), and survival (Harrell's *C*, integrated Brier score IBS, calibration slope). Values are mean, ±, 95% CI.
Notes. *SS64* Siemens Sensation 64 scanner, *PB* Philips Brilliance scanner. "1.4" and "0.9" denote table pitch values (mm per rotation) used in CT acquisition protocols.

Table 10 | CT colonography (ACRIN 6664) detection by acquisition and polyp size

Subset	Precision	Recall	F1
All exams (prone+supine)	0.878 ± 0.015	0.871 ± 0.018	0.874 ± 0.017
Prone only	0.880 ± 0.017	0.882 ± 0.019	0.881 ± 0.018
Supine only	0.883 ± 0.016	0.894 ± 0.018	0.889 ± 0.017
Polyps 6–9 mm	0.835 ± 0.020	0.858 ± 0.022	0.846 ± 0.021
Polyps ≥ 10 mm	0.910 ± 0.014	0.896 ± 0.016	0.903 ± 0.015

Values are mean, ±, 95% CI.

$C = 0.832$ with IBS = 0.161 and a calibration slope of 1.01, outperforming Cox+radiomics and DeepSurv, and exceeding transformer/Mamba baselines while maintaining superior calibration. Ablations clarify complementary roles. Removing AnatoTI reduces boundary accuracy on both modalities; removing APSI primarily lowers detection recall and degrades survival discrimination and calibration. The full configuration yields the strongest survival performance and the lowest prediction error, indicating that anatomy-aware token interpolation and prototype-driven context injection together are necessary to realize the model's performance envelope. The results show that the proposed modules, coupled with a state-space backbone, provide an effective and calibrated multimodal solution for colorectal cancer segmentation, detection, and survival prediction.

The current work has a number of shortcomings. The retrospective public datasets on which all analyses are based have variable annotation quality, small sample sizes for certain tasks (particularly survival prediction), and no paired endoscopy-CT data at the subject level, which limits the evaluation of true cross-modal complementarity. External validation is limited to a small number of centres and acquisition protocols, and comparisons with prior survival models rely on results reported in separate cohorts rather than direct head-to-head evaluation. Prospective, multicenter validation, the addition of other modalities like MRI, pathology, and molecular data, the creation of paired multimodal datasets to better analyze cross-modal interactions, and research into workflow integration and practical clinical impact are all important areas for future work.

Methods

Problem setup and overview

Given a corpus of unpaired endoscopic RGB images $E \in \mathbb{R}^{H_e \times W_e \times 3}$ and pelvic CT volumes $V \in \mathbb{R}^{H_c \times W_c \times D}$, we introduce *HydraMamba* shown in Fig. 6 a unified architecture for unpaired multi-modal learning. The model is designed to produce (i) a tumor segmentation mask M on a reference imaging space (CT slice space by default), (ii) lesion bounding boxes $\mathcal{B} = \{b_k\}$, and (iii) a survival risk representation r used to estimate patient hazard $h(t|r)$. The framework is built upon the principle of representation disentanglement, decomposing inputs into a modality-invariant anatomical representation and a modality-specific style representation. These are processed by a shared selective state space backbone (Mamba)[17] augmented with two modules: *Anatomy Aware Token Interpolation (AnatoTI)*, which performs masked token reconstruction within the shared anatomical space, and *Anatomical Prototype State Injection (APSI)*, which injects global anatomical context derived from the shared representation. A late-fusion mechanism combines the anatomical and style representations to feed task-specific "hydra" heads for segmentation, detection, and survival.

Preprocessing and tokenization

Offline Feature Extraction: All input images are first processed by a frozen, pre-trained MedSigLIP image encoder. For each 2D CT slice or endoscopic frame, the encoder produces a spatial feature map. These feature maps, which represent high-level visual information, are saved and used as the direct input for all subsequent model training and evaluation. This step is performed only once.

Tokenization: The trainable part of our model begins by taking these pre-computed feature maps. Tokens for the Mamba backbone are formed by flattening patches from these feature maps. For CT, we use 2D per-slice patching with slice index encodings. For endoscopy, we use non-overlapping patches from the feature map.

State space encoder and representation disentanglement

We adopt the discretized state space recurrence used in modern Mamba-style vision encoders[17,27]. For input tokens $\{x_i\}_{i=1}^{L}$ with learned, data-dependent time scale Δ_i, the hidden state $h_i \in \mathbb{R}^d$ and output $y_i \in \mathbb{R}^d$ evolve as

$$h_i = \bar{A}_i h_{i-1} + \bar{B}_i x_i, \tag{1}$$

$$y_i = C h_i + D x_i, \tag{2}$$

where

$$\bar{A}_i = \exp(\Delta_i A) \tag{3}$$

$$\bar{B}_i \approx (\Delta A)^{-1}(\exp(\Delta_i A) - I)\Delta_i B \tag{4}$$

For learned A, B, C, D. This causal formulation provides linear time long range modeling but suffers from long range decay when interactions must traverse many steps; alleviating this limitation requires non-causal augmentation of the output mapping C or multi-directional scanning.

Shared Backbone and Disentangled Encoders: To enable learning from unpaired data, we employ a shared backbone architecture that processes inputs from either modality using the same set of Mamba block weights. To handle the significant statistical shifts between endoscopy and CT data, we incorporate modality-specific LayerNorm parameters within the shared backbone. The architecture uses two parallel encoders:

A deep, shared Mamba-based backbone that encodes the input image into a modality-invariant **anatomical representation** s, which captures the underlying anatomical content. A shallow, modality-specific convolutional encoder that produces a compact **modality representation***z*, which encodes the appearance and style characteristics of the source modality (e.g., CT or endoscopy).

This explicit separation of content and style is fundamental to preventing information leakage and enabling robust unpaired learning.

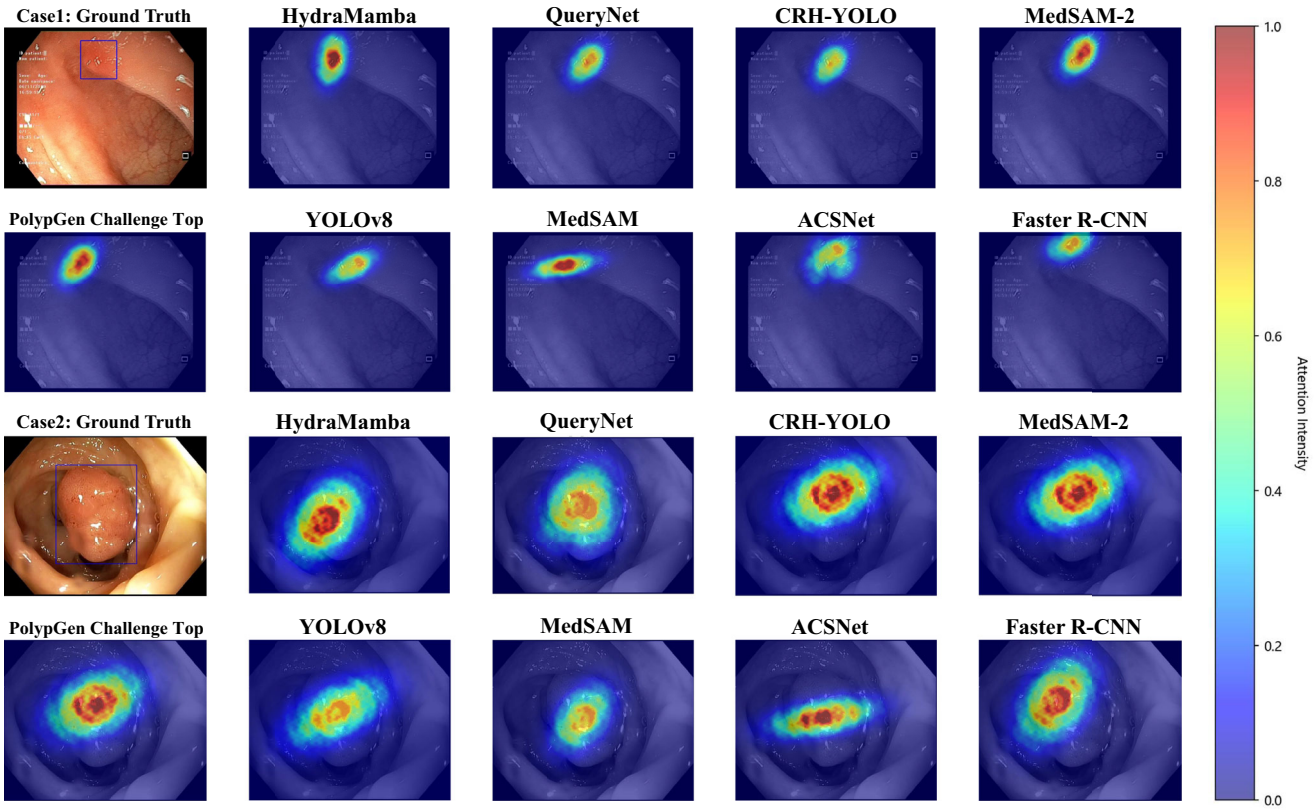


Fig. 5 | Qualitative comparison of attention maps for lesion detection on the endoscopy dataset. For two representative cases (Case 1 and Case 2), we visualize the ground truth bounding box and the corresponding attention heatmaps for HydraMamba and various baseline models. Red/yellow indicates high-attention regions, while blue indicates low attention. HydraMamba's attention is clearly and accurately localized on the lesion, demonstrating superior focus compared to the compared baselines.

Tokenization and Scanning: Endoscopy is embedded by a 3×3 stride-2 convolutional stem followed by non-overlapping $P \times P$ patch flattening and positional encodings. CT is embedded with either (i) a $3 \times 3 \times 3$ stem and $P \times P \times S$ cuboid tokens and (ii) 2D per slice patching with slice index encodings. For a given input, a scan π (Hilbert or Morton) defines a 1D token order that preserves local adjacency.

Anatomy aware token interpolation (AnatoTI)

Motivation: Replacing missing or low-confidence tokens with a generic learnable vector disrupts the causal, input dependent selective scan property of SSMs and degrades representation quality. Mamba specific masked modeling succeeds when masked tokens are generated via state-space consistent interpolation rather than free parameters. Our AnatoTI module reinforces the learning of a robust, modality-invariant anatomical representation. It operates on the anatomical representations to perform masked token reconstruction. This acts as a self-supervised objective that encourages the shared backbone to learn universal anatomical features, consistent with the causal dynamics of SSMs, without being influenced by modality-specific shown in Fig. 7.

Geometry and uncertainty aware interpolation: Let π denote the scan over the anatomical tokens S, and suppose indices $i < j$ are valid tokens with a gap $G(i, j) = i + 1, \ldots, j - 1$. For $\alpha = 1, \ldots, j - i - 1$, define $u = \alpha/(j - i)$ and construct:

$$\widehat{s}_{i+\alpha} = \gamma_{i+\alpha}\, \overrightarrow{s_{i+\alpha}} + (1 - \gamma_{i+\alpha}) \overleftarrow{s_{i+\alpha}}, \tag{5}$$

$$\overrightarrow{s_{i+\alpha}} = w_\alpha\, s_i + (1 - w_\alpha)\, \phi_\theta(\mathcal{N}(i),\, u), \tag{6}$$

$$w_\alpha = \sigma(\beta(1 - u)), \tag{7}$$

$$\overleftarrow{s_{i+\alpha}} = \widetilde{w}_\alpha\, s_j + (1 - \widetilde{w}_\alpha)\, \phi_\theta(\mathcal{N}(j),\, 1 - u), \tag{8}$$

$$\widetilde{w}_\alpha = \sigma(\beta u). \tag{9}$$

where ϕ_θ maps local image descriptors in neighborhoods $\mathcal{N}_i, \mathcal{N}_j$ (e.g., color gradients for endoscopy; HU histograms for CT) and the geodesic progress u to a token in $\mathbb{R}^d$. The blend gate $\gamma_{i+\alpha} = \sigma(\eta - \kappa \varsigma_{i+\alpha})$ is uncertainty-aware, where $\varsigma_{i+\alpha}$ estimates aleatoric uncertainty.

Causality at deployment via dual pass training: During training we use a symmetric, *teacher forcing* blend in (5). At inference, to preserve strict causality, we freeze $\gamma_{i+\alpha} = 1$ and disable the backward term, yielding a purely forward interpolation that depends only on past tokens.

Where AnatoTI applies: We invoke AnatoTI during pretraining-style consistency regularization by replacing a fraction of anatomical tokens s_i with masked placeholders and training the model to reconstruct them.

Anatomical prototype state injection (APSI)

Motivation: The causal SSM in (2) constrains each token to "see" only its predecessors along π, which limits global perception. Non-causal augmentation that modulates the output pathway enables one-pass global context without multi-directional scans. The core mechanism derives patient-specific anatomical prototypes from the shared content space S and injects this context via a structured, token-conditioned low-rank update. Modality-specific style information is fused at a late stage to inform the final task-specific predictions shown in Fig. 8.

Prototype construction: Given the anatomical representation tokens $\{S_i\}_{i=1}^{L}$ (optionally whitened), we learn K prototypes via temperature-controlled soft clustering:

$$a_i = W_a s_i,\; c_{ik} = \mathrm{softmax}k(a_i^\top q_k/\tau), \tag{10}$$

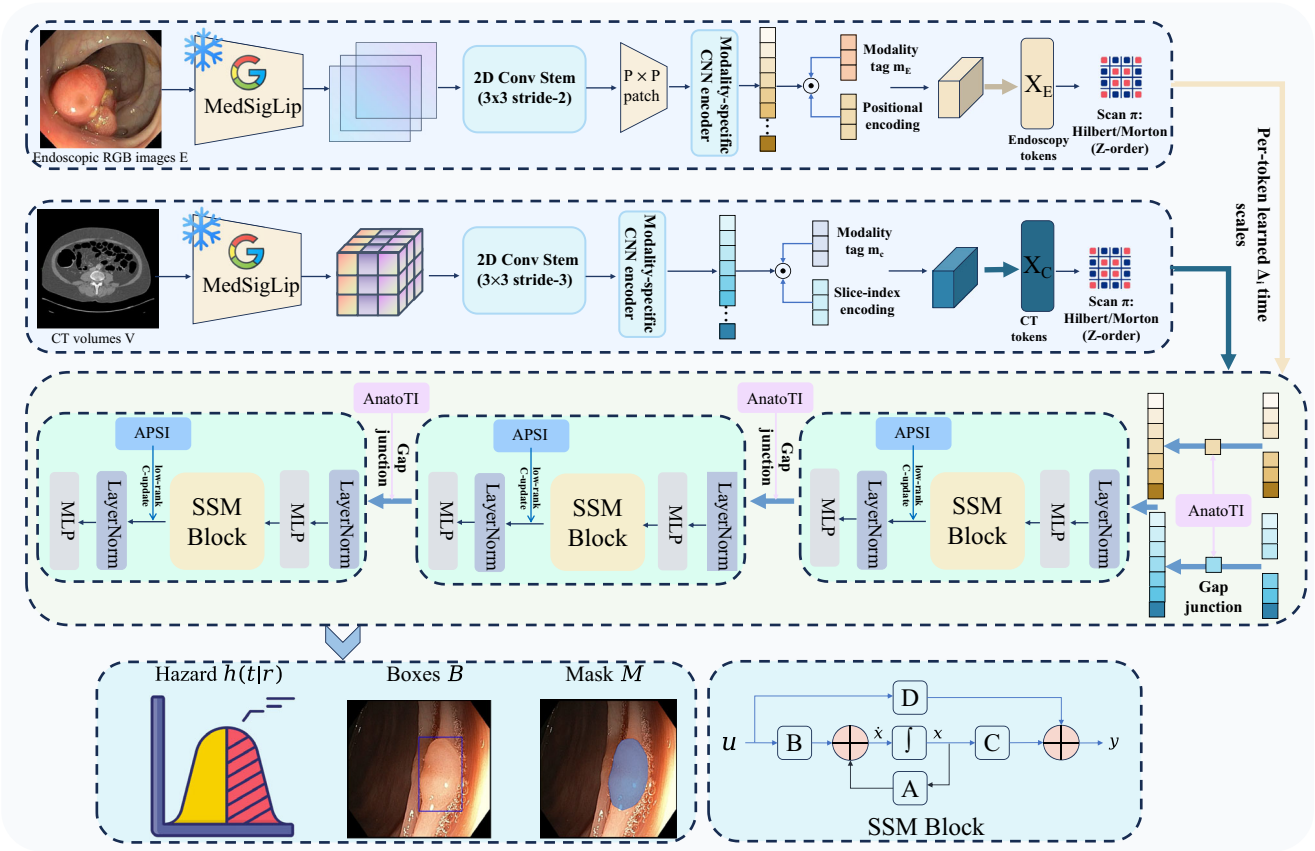


Fig. 6 | The HydraMamba multimodal architecture. The end-to-end framework for comprehensive colorectal cancer analysis.

$$p_k = \frac{\sum_i c_{ik} s_i}{\sum_i c_{ik}},\ k = 1, \ldots, K, \tag{11}$$

with trainable queries $\{q_k\}$ and projection W_a. We encourage diversity via an entropy regularizer $-\sum_{i,k} c_{ik} \log c_{ik}$. Low rank operator injection and Late Fusion: For token i, we form an anatomical context vector $g_i = \sum_k c_{ik} p_k$ and modulate the SSM output mapping by a rank$-r$ update on the anatomical weights:

$$h_i = \bar{A}_i h_{i-1} + \bar{B}_i s_i, \tag{12}$$

$$y_i^s = (C + U \mathrm{diag}(\Gamma g_i) V^\top) h_i + D s_i, \tag{13}$$

where y_i^s is the context-aware anatomical representation. The final representation for the task heads, y_i^{final}, is produced by a late-fusion module (a gated MLP) that combines the enhanced anatomical features y_i^s with the modality-style vector z : $y_i^{\text{final}} = Fusion(y_i^s, z)$. This ensures that the deep backbone learns a pure anatomical representation, while allowing modality-specific characteristics to inform the final prediction.

Stability refinements: We apply (i) a nuclear norm penalty $\lambda^* \sum_i |U \mathrm{diag}(\Gamma g_i) V^\top|_*$ to prevent rank inflation; (ii) *context drop* with small probability to avoid over-reliance on APSI; and (iii) prototype EMA updates $p_k \leftarrow \rho p_k + (1-\rho)\widehat{p}_k$ for smooth dynamics.

Backbone stack and multi-scale features

We interleave shared SSM blocks with depthwise separable convolutions for local mixing. Each stage uses modality-specific LayerNorm → SSM (with AnatoTI at inputs) → modality-specific LayerNorm → pointwise MLP, with residual connections. The APSI is applied in every other block to the anatomical representations. We build a feature pyramid by down sampling token density across stages, exposing high resolution features to segmentation and mid/low resolution features to detection and survival[18].

Task-specific decoders

Segmentation: A U-shaped decoder upsamples multi-scale SSM features to the CT slice space. Each up block performs 2× upsampling, depthwise convolution, and gated MLP fusion with lateral skips from early stems. The final 1×1 projection yields logits ℓ_{seg}; loss combines soft Dice and focal cross entropy. To sharpen boundaries, we add a surface loss on the signed distance transform and a topology-preserving penalty on Euler characteristic in a narrow band.

Detection (bounding boxes): an anchor-free, center-based head operates on a 1/4–1/8 scale feature map and predicts objectness $o(u, v)$ and box parameters (l, t, r, b) with centerness weighting. Training uses focal loss for classification and a mixture of IoU/GIoU and smooth-ℓ_1 for regression.

Survival: We form a patient embedding r via attention pooling concentrated on predicted tumor regions: tokens receive attention weights $\alpha_i \propto \exp(w^\top [y_i, \widehat{m}_i])$ where $\widehat{m}_i$ are upsampled mask logits. We fit either (i) a Cox head that outputs log risk $\phi = w_\phi^\top r$ optimized by the negative partial log likelihood, or (ii) a discrete time head that outputs hazards

$$\lambda_t = \sigma(w_t^\top r) \tag{14}$$

over T bins with likelihood

$$\mathcal{L} = \prod_p \prod_{t \le T} \left(1 - \lambda_t^{(p)}\right)^{\mathbb{1}[T_p > t]} \left(\lambda_t^{(p)}\right)^{\mathbb{1}[T_p = t, \delta_p = 1]} \tag{15}$$

To encourage clinically meaningful cues, a weak radiomics prior u (mask area, elongation, detector counts) is fused via $r \leftarrow r + W_u u$.

Fig. 7 | AnatoTI construction.

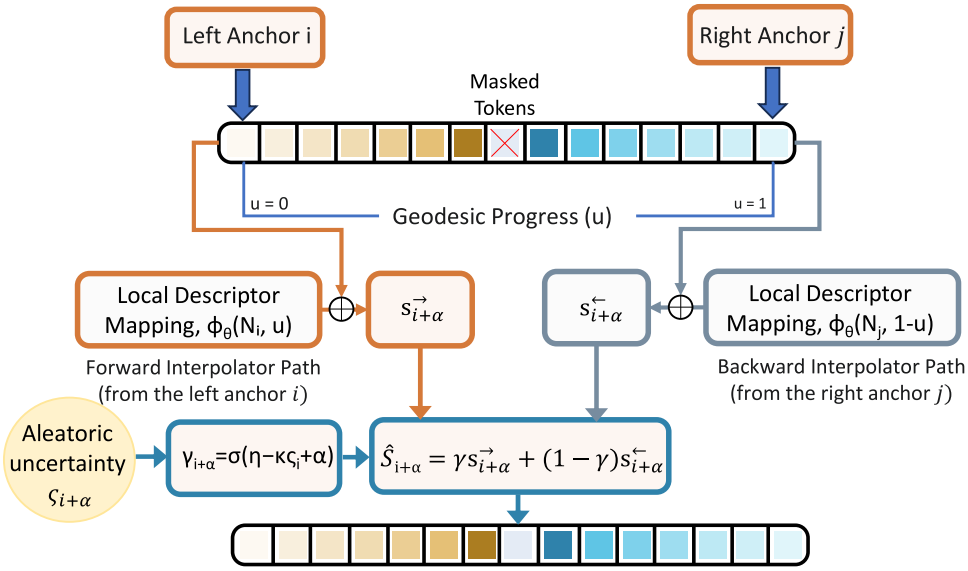
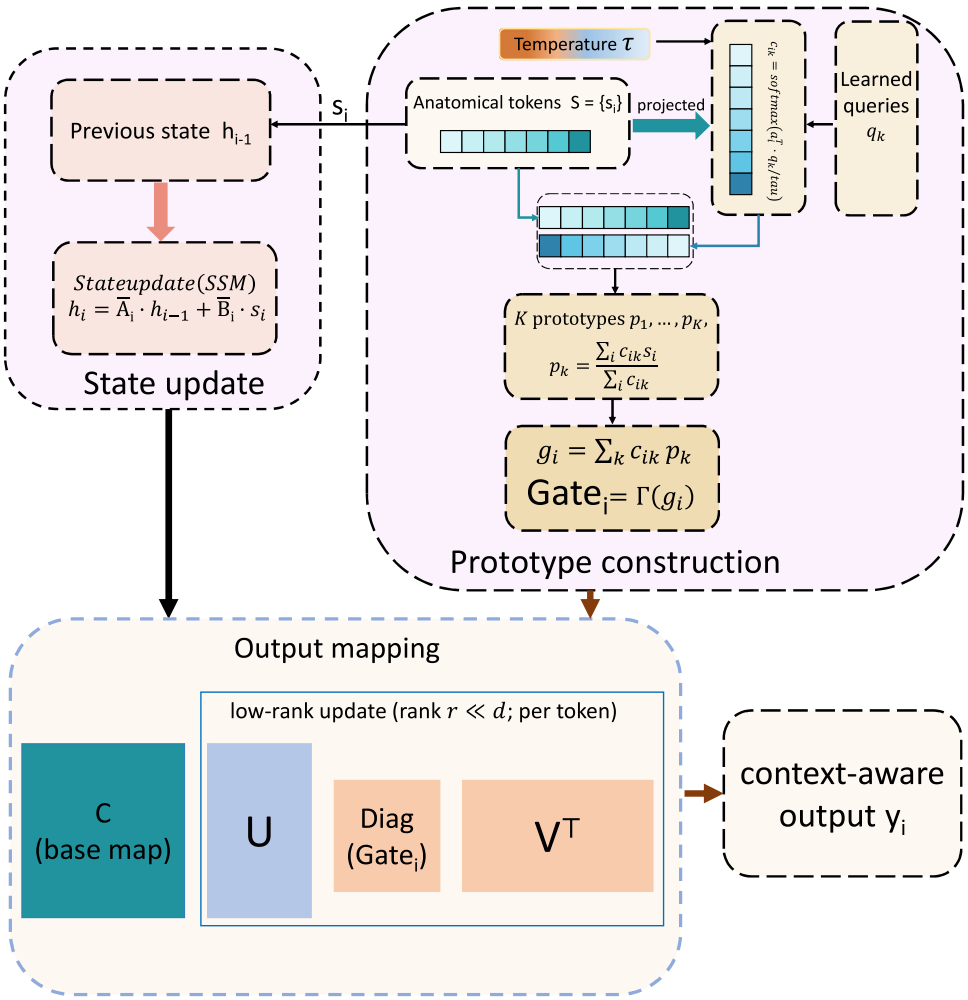


Fig. 8 | APSI construction.



Training objective and optimization

The total loss is

$$\mathcal{L} = \lambda_{\text{task}}\mathcal{L}_{\text{task}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{dis}}\mathcal{L}_{\text{dis}} + \lambda_{\mathcal{R}}\mathcal{R} + \lambda_{\text{wd}}\|\Theta\|_2^2 \quad (16)$$

Disentanglement Loss ($\mathcal{L}dis$): To enforce a clean separation of anatomical (s) and modality (z) representations, we use a margin-based similarity loss. It ensures that anatomical representations of similar structures are close regardless of modality, while modality representations are close for the same modality regardless of anatomical content. For subjects p, q, and modalities i, j:

$$\begin{aligned}\mathcal{L}_{dis} = \; & \mathbb{E}[\max(0, \alpha_s - \cos(s_p^i, s_p^j) + \cos(s_p^i, s_q^i))] \\ & + \mathbb{E}[\max(0, \alpha_z - \cos(z_p^i, z_q^i) + \cos(z_p^i, z_p^j))]\end{aligned} \quad (17)$$

Semantic Alignment Loss ($\mathcal{L}align$): To align the anatomical spaces of the two modalities, we use a knowledge-guided contrastive loss inspired by MedCLIP.[5] For unpaired batches from each modality, we construct a semantic similarity matrix S based on weak labels (presence of tumor, organ labels). The loss then minimizes the divergence between the computed cross-modal representation similarity and S, encouraging anatomically similar (but unpaired) images to have similar representations. Task-Specific Loss ($\mathcal{L}task$): The standard supervised losses for segmentation, detection, and survival, applied to samples with available ground-truth labels.

Training protocol and hyperparameters

We use patient-level k-fold cross-validation on the CT cohort. In each fold, the model is trained on the CT training set features combined with all available endoscopy features. Evaluation is performed on the held-out CT test set features.

The entire HydraMamba model, including the shared Mamba backbone, AnatoTI, APSI, and task-specific heads, is trained from scratch on the pre-extracted MedSigLIP features. We use the AdamW optimizer with a cosine learning rate decay and mixed precision. The total loss combines task-specific losses with the cross-modal disentanglement and alignment losses that enforce a coherent shared representation space.

To eliminate any risk of information leakage, all preprocessing and any statistic-fitting operations were confined to the training data of each outer fold. Feature extraction with the MedSigLIP encoder is a fixed, frozen mapping from pixels to features and does not learn from our data; nevertheless, any downstream transformations that require data-driven parameters were estimated on the training partition only and then applied unchanged to validation/test partitions. Splits were enforced at the patient level across all tasks so that a subject never appears in more than one of train/validation/test, and the survival head consumed only CT patient embeddings with linked outcomes.

Survival modeling used nested, patient-level resampling. In the outer loop we performed $k = 5$-fold cross-validation with folds stratified by event indicator and approximate follow-up quantiles. Within each outer training split, 20% of patients were held out as an inner validation set for model selection and early stopping; the inner set was never used to compute the reported test metrics. Hyperparameters were tuned strictly inside this inner loop (learning rate/weight decay/dropout and APSI design), using a grid that included $K \in 16, 32, 64$ prototypes and ranks $r \in 4, 8, 16$. The configuration that maximized Uno's $C(t)$ at 1 year while minimizing IBS on the inner validation set was then refit on the full outer-training fold and evaluated once on the untouched outer-test fold. Uncertainty was quantified with patient-level nonparametric bootstrapping. After aggregating out-of-fold predictions from the five outer test folds, we computed 95% confidence intervals for Harrell's C, Uno's $C(t)$ at 1 year, the integrated Brier score (IBS), and the calibration slope using $B = 1000$ bootstrap replicates, resampling patients (keeping all of a patient's slices/frames together) to preserve within-subject dependence.

Ethics and consent to participate

This study was conducted using publicly available, fully de-identified datasets obtained from open-access repositories (TCGA-COAD, TCGA-READ, StageII-Colorectal-CT, PolypGen, CVC-ColonDB, and ACRIN 6664). As no human participants or animals were directly involved and all data were previously anonymized, ethics approval and consent to participate were not required.

Data availability

All datasets used in this study are publicly accessible:• StageII-Colorectal-CT—Abdominal or pelvic enhanced CT images within 10 days before surgery of 230 patients with stage II colorectal cancer: https://www.cancerimagingarchive.net/collection/stageii-colorectal-ct/• PolypGen: https://www.synapse.org/Synapse:syn26376615/wiki/613312• CVC-colonDB: https://vi.cvc.uab.es/colon-qa/cvccolondb/• CT COLONOGRAPHY (ACRIN 6664): https://www.cancerimagingarchive.net/collection/ct-colonography/• TCGA-READ: https://www.cancerimagingarchive.net/collection/tcga-read/• TCGA-COAD: https://www.cancerimagingarchive.net/collection/tcga-coad/. The custom scripts developed for data pre-processing, model training and evaluation will be made publicly available on GitHub upon publication. These scripts are implemented on standard PyTorch frameworks and can be used to fully reproduce all analyses.

Code availability

The custom scripts developed for data pre-processing, model training and evaluation will be made publicly available on GitHub upon publication. These scripts are implemented on standard PyTorch frameworks and can be used to fully reproduce all analyses.

Received: 16 October 2025; Accepted: 30 November 2025;
Published online: 31 December 2025

References

1. Qi, L.-Y. et al. Research status and trends of deep learning in colorectal cancer (2011-2023): Bibliometric analysis and visualization. *World J. Gastrointest. Oncol.* **17**, 103667 (2025).
2. Dimitriou, N., Arandjelović, O., Harrison, D. J. & Caie, P. D. A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis. *npj Digit. Med.* **1**, 52 (2018).
3. Zhang, C. et al. Development and validation of an AI-based multimodal model for pathological staging of gastric cancer using CT and endoscopic images. *Acad. Radiol.* **32**, 2604–2617 (2025).
4. Tian, R. et al. Multimodal fusion model for prognostic prediction and radiotherapy response assessment in head and neck squamous cell carcinoma. *npj Digit. Med.* **8**, 302 (2025).
5. Li, H. et al. Systematic review and meta-analysis of deep learning for MSI-H in colorectal cancer whole slide images. *npj Digit. Med.* **8**, 456 (2025).
6. Wulczyn, E. et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digit. Med.* **4**, 71 (2021).
7. Ali, S. et al. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. *Sci. Rep.* **14**, 2032 (2024).
8. Chen, Z. et al. Masked image modeling advances 3d medical image analysis. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 1970–1980 (IEEE, 2023).
9. Jiang, Y. et al. Biology-guided deep learning predicts prognosis and cancer immunotherapy response. *Nat. Commun.* **14**, 5135 (2023).
10. Zhang, M. et al. An interpretable ct-based deep learning model for predicting overall survival in patients with bladder cancer: a multicenter study. *npj Precis. Oncol.* **9**, 288 (2025).
11. Wang, Y.-Y., Liu, B. & Wang, J.-H. Application of deep learning-based convolutional neural networks in gastrointestinal disease endoscopic examination. *World J. Gastroenterol.* **31**, 111137 (2025).

12. Osco, L. P. et al. The segment anything model (SAM) for remote sensing applications: from zero to one shot. *Int. J. Appl. Earth Observ. Geoinf.* **124**, 103540 (2023).
13. Ma, J. et al. Medsam2: segment anything in 3d medical images and videos. *arXiv* preprint https://doi.org/10.48550/arXiv.2504.03600 (2025).
14. Vaswani, A. et al. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems* **30** https://doi.org/10.48550/arXiv.1706.03762 (2017).
15. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint* https://doi.org/10.48550/arXiv.2010.11929 (2020).
16. Somvanshi S, et al. From S4 to Mamba: A Comprehensive Survey on Structured State Space Models. arXiv preprint arXiv:2503.18970, 2025.
17. Gu, A. & Dao, T. Mamba: linear-time sequence modeling with selective state spaces https://arxiv.org/abs/2312.00752 (2024).
18. Tang, F. et al. Mambamim: pre-training mamba with state space token interpolation and its application to medical image segmentation. *Med. Image Anal.* **103**, 103606 (2025).
19. Ali S, et al. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Sci. Data* **10**, 75 (2023).
20. Bernal, J., Sánchez, J. & Vilarino, F. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognit.* **45**, 3166–3182 (2012).
21. Tong, T. & Li, M. Abdominal or pelvic enhanced CT images within 10 days before surgery of 230 patients with stage II colorectal cancer (StageII-Colorectal-CT) [Dataset]. The Cancer Imaging Archive. https://doi.org/10.7937/p5k5-tg43 (2022).
22. Smith, K. et al. Data From CT COLONOGRAPHY (ACRIN 6664) (Version 1) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.NWTESAY1 (2015).
23. Kirk, S., Lee, Y., Sadow, C. A., & Levine, S. The Cancer Genome Atlas Rectum Adenocarcinoma Collection (TCGA-READ) (Version 3) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.F7PPNPNU (2016).
24. Kirk, S. et al. The Cancer Genome Atlas Colon Adenocarcinoma Collection (TCGA-COAD) (Version 3) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.HJJHBOXZ (2016).
25. Katzman, J. L. et al. Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 24 (2018).
26. Jiang, X. et al. An MRI deep learning model predicts outcome in rectal cancer. *Radiology* **307**, e222223 (2023).
27. Zhu L, et al. Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417, 2024.
28. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation https://arxiv.org/abs/1802.02611 (2018).
29. Fan, D.-P. et al. Pranet: parallel reverse attention network for polyp segmentation. In *Proc. International Conference on Medical Image Computing and Computer-assisted Intervention*, 263–273 (Springer, 2020).
30. Duc N. T. et al. Colonformer: An efficient transformer based method for colon polyp segmentation. IEEE Access, 2022, 10: 80575-80586.
31. Nachmani, R., Nidal, I., Robinson, D., Yassin, M. & Abookasis, D. Segmentation of polyps based on pyramid vision transformers and residual block for real-time endoscopy imaging. *J. Pathol. Inform.* **14**, 100197 (2023).
32. Liu, D., Lu, C., Sun, H. & Gao, S. Na-segformer: a multi-level transformer model based on neighborhood attention for colonoscopic polyp segmentation. *Sci. Rep.* **14**, 22527 (2024).
33. Hottung A., Mahajan M. & Tierney K. PolyNet: Learning diverse solution strategies for neural combinatorial optimization. arXiv preprint arXiv:2402.14048 (2024).
34. Nguyen, Q. V., Vo, T. H. S., Kang, S.-R. & Kim, S.-H. Polyp-ses: automatic polyp segmentation with self-enriched semantic model. In *Proc. Asian Conference on Computer Vision*, 2803–2819 (Springer, 2024).
35. Shao, R., Bi, X.-J. & Chen, Z. Hybrid vit-cnn network for fine-grained image classification. *IEEE Signal. Process. Lett.* **31**, 1109–1113 (2024).
36. Cai, L., Chen, L., Huang, J., Wang, Y. & Zhang, Y. Know your orientation: a viewpoint-aware framework for polyp segmentation. *Med. Image Anal.* **97**, 103288 (2024).
37. Hu, B. C. et al. PraNet-V2: Dual-Supervised Reverse Attention for Medical Image Segmentation. arXiv preprint arXiv:2504.10986, (2025).
38. Xie, J. et al. Promamba: Prompt-mamba for polyp segmentation. arXiv preprint arXiv:2403.13660 (2024).
39. Archit A. Vim-unet: Vision mamba for biomedical segmentation. arXiv preprint arXiv:2404.07705 (2024).
40. Ren, S. et al. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 39, 1137–1149 (2016).
41. Zhang, R. et al. Adaptive context selection for polyp segmentation https://arxiv.org/abs/2301.04799 (2023).
42. Khan, Z. F. et al. Real-time polyp detection from endoscopic images using YOLOv8 with YOLO-score metrics for enhanced suitability assessment. In *Proc. IEEE Access* (IEEE, 2024).
43. Wan, J. et al. Crh-yolo for precise and efficient detection of gastrointestinal polyps. *Sci. Rep.* **14**, 30033 (2024).
44. Ali, S. et al. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Sci. Data* **10**, 75 (2023).
45. Lei M. et al. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. arXiv preprint arXiv:2506.17733, 2025.
46. Ma J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
47. Chai, J. et al. Querynet: a unified framework for accurate polyp segmentation and detection. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* 544–554 (Springer, 2024).
48. Zhu J, et al. Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874, 2024.
49. Oktay O, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999, 2018.
50. Zhang, M. et al. Dynamic context-sensitive filtering network for video salient object detection. In *Proc. IEEE/CVF International Conference on Computer Vision* 1553–1563 (IEEE, 2021).
51. Ding, A. S. et al. A self-configuring deep learning network for segmentation of temporal bone anatomy in cone-beam CT imaging. *Otolaryngol. Head. Neck Surg.* **169**, 988–998 (2023).
52. Wang, Y. et al. Deep distance transform for tubular structure segmentation in CT scans. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3833–3842 (IEEE, 2020).
53. Tang, Y. et al. Self-supervised pre-training of Swin Transformers for 3d medical image analysis. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20730–20740 (IEEE, 2022).
54. Lee, H. et al. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. arXiv preprint arXiv:2209.15076 (2022).
55. Cheng, X. et al. Implicit motion handling for video camouflaged object detection. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13864–13873 (IEEE, 2022).
56. Zhang, R. et al. Ag-crc: Anatomy-guided colorectal cancer segmentation in ct with imperfect anatomical knowledge. arXiv preprint arXiv:2310.04677 (2023).
57. Lin, J. et al. Shifting more attention to breast lesion segmentation in ultrasound videos. In *Proc. International Conference on Medical*

Image Computing and Computer-Assisted Intervention 497–507 (Springer, 2023).
58. Yao, L. et al. A colorectal coordinate-driven method for colorectum and colorectal cancer segmentation in conventional ct scans. *IEEE Trans. Neural Netw. Learn. Syst.* **36**, 7395–7406 (2024).
59. Xing, Z., Ye, T., Yang, Y., Liu, G. & Zhu, L. Segmamba: long-range sequential modeling mamba for 3d medical image segmentation https://arxiv.org/abs/2401.13560 (2024).
60. Raju, A. S. N. et al. A hybrid framework for colorectal cancer detection and U-Net segmentation using polynetdwtcadx. *Sci. Rep.* **15**, 847 (2025).
61. Khalid, M., Deivasigamani, S., V, S. & Rajendran, S. An efficient colorectal cancer detection network using atrous convolution with coordinate attention transformer and histopathological images. *Sci. Rep.* **14**, 19109 (2024).
62. Guo, W. et al. Meply: a large-scale dataset and baseline evaluations for metastatic perirectal lymph node detection and segmentation https://arxiv.org/abs/2404.08916 (2024).
63. Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
64. Chen, M., Wang, K. & Wang, J. Vision transformer-based multilabel survival prediction for oropharynx cancer after radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **118**, 1123–1134 (2024).
65. Wang, Y., Huang, N., Li, T., Yan, Y. & Zhang, X. Medformer: a multi-granularity patching transformer for medical time-series classification. *Adv. Neural Inf. Process. Syst.* **37**, 36314–36341 (2024).
66. Ding, K., Zhou, M., Metaxas, D. N. & Zhang, S. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* 622–631 (Springer, 2023).
67. Gomaa, A. et al. Comprehensive multimodal deep learning survival prediction enabled by a transformer architecture: a multicenter study in glioblastoma. *Neurooncol. Adv.* **6**, 122 (2024).
68. Xia, Y. et al. CT-based multimodal deep learning for non-invasive overall survival prediction in advanced hepatocellular carcinoma patients treated with immunotherapy. *Insights Into Imag.* **15**, 214 (2024).
69. Xu, Y. & Chen, H. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proc. IEEE/CVF International Conference on Computer Vision* 21241–21251 (IEEE, 2023).
70. Zhang, J. et al. 2dmamba: efficient state space model for image representation with applications on giga-pixel whole slide image classification. In *Proc. Computer Vision and Pattern Recognition Conference* 3583–3592 (Computer Vision Foundations, 2025).
71. Zhang, C. et al. ME-Mamba: Multi-Expert Mamba with Efficient Knowledge Capture and Fusion for Multimodal Survival Analysis. arXiv preprint arXiv:2509.16900 (2025).

Acknowledgements

The authors would like to acknowledge that there are no additional contributions to declare. This study was supported by the Joint Funds for the Innovation of Science and Technology, Fujian Province (Grant number: 2023Y9299, to Chenshen Huang).

Author contributions

Conceptualization, J.L., and C.H.; Methodology, Y.J.; Literature research, N.W., W.L., Y.L., Z.N., and Q.H.; Data acquisition, Y.J., H.C., and C.H.; Data analysis and interpretation, N.W., Y.J., and C.H.; Writing—original draft, N.W., J.L., and Q.Y.; Writing—review and editing, W.L., Y.L., H.C., and C.H.; Funding acquisition: C.H.; All authors read and approved the submitted version of manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Hong Chen, Qiang Yan or Chenshen Huang.