

MRICombo: a deep-learning-based framework for universal volume segmentation, grading-staging, and malignancy detection across heterogeneous MRI

Tao Tan

taotan@mpu.edu.mo

Faculty of Applied Sciences, Macao Polytechnic University

Zhuoneng Zhang

Faculty of Applied Sciences, Macao Polytechnic University

Luyi Han

Netherlands Cancer Institute

Dengqiang Jia

Macao Polytechnic University

Chenggong Yan

Department of Medical Imaging Center, Nanfang Hospital, Southern Medical University

Tianyu Zhang

Radboud University Medical Centre <https://orcid.org/0009-0007-4724-719X>

Jiaju Huang

Macao Polytechnic University

Da Huang

Macao Polytechnic University <https://orcid.org/0000-0002-6375-0197>

Yapeng Wang

Faculty of Applied Sciences, Macao Polytechnic University

Shandong Wu

University of Pittsburgh <https://orcid.org/0000-0002-0770-2203>

Michiel Kallenberg

Artificial Intelligence Group, Wageningen University & Research, The Netherlands

<https://orcid.org/0000-0002-4661-6674>

Nico Karssemeijer

Radboud University Medical Center

Ritse Mann

Radboud UMC <https://orcid.org/0000-0001-8111-1930>

Yue Sun

Article

Keywords: MRI, Heterogeneous sequences, Multi-expert framework, Universal Volume Segmentation, Grading, Staging, and Malignancy Detection

Posted Date: November 25th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-8096445/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

MRICombo: a deep-learning-based framework for universal volume segmentation, grading-staging, and malignancy detection across heterogeneous MRI

Zhuoneng Zhang¹, Luyi Han^{2,3}, Dengqiang Jia¹, Chenggong Yan⁷, Tianyu Zhang^{2,3}, Jiaju Huang¹, Da Huang¹, Yapeng Wang¹, Shandong Wu⁴, Michiel Kallenberg⁵, Nico Karssemeijer^{3,6}, Ritse Mann^{2,3}, Yue Sun^{1,*}, and Tao Tan^{1,*}

¹Faculty of Applied Sciences, Macao Polytechnic University, Rua de Luis Gonzaga Gomes, Macao, China

²Netherlands Cancer Institute, Department of Radiology, Amsterdam, 1066 CX, The Netherlands

³Radboud University Medical Centre, Department of Radiology and Nuclear Medicine, Nijmegen, 6525 GA, The Netherlands

⁴Departments of Radiology, Biomedical Informatics, Bioengineering, and Computer Science, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260, USA

⁵Artificial Intelligence Group, Wageningen University & Research, The Netherlands

⁶ScreenPoint Medical BV, Nijmegen, The Netherlands.

⁷Department of Medical Imaging Center, Nanfang Hospital, Southern Medical University, Guangzhou, China

*corresponding author(s): yuesun@mpu.edu.mo, taotan@mpu.edu.mo

ABSTRACT

Comprehensive magnetic resonance imaging (MRI) analysis in oncology involves multiple interrelated tasks including volumetric segmentation, grading, staging, and malignancy detection. However, most existing deep learning models are task-specific or sequence-specific, lacking the generalizability required for heterogeneous sequences. Here we present MRICombo, a unified multi-expert deep learning framework for universal anatomical delineation and tumor characterization across 9 heterogeneous imaging sequences. Developed using 7,376 MRI sequences from 2,282 individuals, MRICombo achieves state-of-the-art performance with mean Dice similarity coefficients of 0.839 for segmenting 14 critical anatomical structures and 0.633 for labeling 11 major tumor types. It also attains a mean area under the receiver operating characteristic curve (AUROC) of 0.913 for glioma grading, bladder and nasopharyngeal cancer staging, and breast and liver tumor malignancy detection. External validation on four independent datasets (998 sequences from 734 individuals) and transfer learning evaluation (260 individuals) confirm robust cross-protocols generalizability. Furthermore, MRICombo supports flexible inference with missing sequences and offers decision interpretability through sequence clustering and expert contribution analysis. As an all-in-one clinical solution, MRICombo significantly reduces deployment costs and streamlines diagnostic workflows, supporting more personalized oncology care.

Keywords: MRI. Heterogeneous sequences. Multi-expert framework. Universal Volume Segmentation, Grading, Staging, and Malignancy Detection.

Introduction

Magnetic Resonance Imaging (MRI) has been as an indispensable clinical diagnostic tool owing to its superior soft tissue contrast and absence of ionizing radiation¹. Multi-sequence MRI is widely employed in disease diagnosis, as different MRI sequences provide complementary information regarding the structural and functional details of organs². For instance, T1-weighted (T1) sequences are effective in visualizing anatomical structures with excellent spatial resolution, while T2-weighted (T2) and Fluid Attenuated Inversion Recovery (FLAIR) sequences exhibit high sensitivity to pathological changes such as edematous tissues. The integration of these diverse MRI sequences yields comprehensive radiological data that significantly enhances diagnostic accuracy and clinical decision-making³.

In clinical practice, multi-sequence MRI enables radiologists to routinely perform complex tasks. For example, they can distinguish malignant tumors from benign ones based on distinct imaging patterns and integrate information from multiple sequences to accurately recognize the boundaries of target regions. However, these manual procedures are time-consuming and exhibits limited reproducibility due to inter-observer variability and subjective interpretation. Specifically, malignant tumor differentiation and tumor grading traditionally require invasive histopathological examination as the "gold standard," which

increases both the complexity and duration of diagnostic workflows and may delay the initiation of optimal treatment. In contrast, deep learning-based methods have demonstrated revolutionary technological advantages in multi-modality image analysis, offering not only automated efficiency but also excellent accuracy and consistency. For instance, the Medical Segmentation Decathlon benchmark has demonstrated that deep learning models can achieve near-expert performance in multi-organ/tumor segmentation. The tasks in this benchmark include the segmentation of the heart, prostate, liver, and pancreas, among others⁴. In tumor classification, another deep learning-based method achieves 91.0% accuracy in benign and malignant differentiation surpassing a traditional method, which yields an accuracy of only 86.4%. Consistent breakthroughs have been reported in similar tasks such as brain tumor grading^{5,6} and breast cancer detection⁷.

Despite these advances, current deep learning-based segmentation or classification models exhibit fundamental limitations that restrict clinical translation. Most segmentation models design dedicated modules for specific organs, e.g., liver⁸, kidney⁹, pancreas¹⁰ or specific tumor types (e.g., glioma¹¹, hepatocellular carcinoma¹², colorectal cancer¹³). While these models perform excellently on specific tasks, they lack cross-organ and cross-disease generalization capabilities. In classification tasks, existing research primarily focuses on benign-malignant differentiation for specific anatomical sites (e.g., breast masses¹⁴, liver lesions¹⁵) or grading/typing tasks for specific tumors, with relatively narrow model applicability. In clinical applications, the deployment of such models typically follows a "one-task-one-model" development workflow, which increases demands on computational resource, storage, and maintenance. It poses significant challenges for adapting to the diverse and evolving requirements of clinical diagnosis. For example, breast tumor diagnosis requires concurrent tumor segmentation¹⁶, classification/grading¹⁷, and prognostic assessment, all of which single-task models cannot accommodate. More importantly, existing deep learning methods predominantly focus on single-task objectives (segmentation or classification), whereas in clinical diagnostic practice, radiologists typically need to simultaneously perform multiple interrelated analytical tasks. Therefore, there is an urgent need for a unified multi-task framework that integrates multiple interrelated analytical tasks such as segmentation and classification.

To address these challenges, medical image analysis is evolving toward generalization and multi-task approaches. In segmentation, two-dimensional universal segmentation models such as MedSAM¹⁸ and SAM-Med2D¹⁹ have demonstrated exceptional performance, but universal segmentation of three-dimensional volumetric images remains more complex due to high annotation costs and computational resource intensity. Although three-dimensional universal segmentation models like SAM-Med3D²⁰ have been developed, they often require multiple interactions to achieve optimal segmentation. However, this is particularly true for complex anatomical structures, and these models lack the efficiency necessary for large-scale automated analysis. In classification, self-supervised learning (SSL) provides a pathway to universal models by leveraging large unlabeled datasets to learn powerful representations, thus diminishing the need for manual annotation. However, prevailing SSL-based strategies²¹⁻²⁴ remain constrained by a two-stage paradigm of pre-training followed by task-specific fine-tuning. This paradigm not only results in a discernible performance gap between stages but also scales inefficiently. Since each downstream task necessitates a separately fine-tuned model, the deployment footprint increases in a manner directly proportional to the number of tasks, thereby failing to deliver a unified "one-model-multi-task" solution. Existing multi-task work^{9,15,25-29}, while capable of simultaneously handling segmentation and classification, is limited to single anatomical site task combinations, unable to meet the complex diagnostic scenario requirements of multi-organ and multi-disease coexistence in clinical practice. More problematically, missing sequence issues are prevalent in clinical practice. Brain tumor diagnosis typically requires complete T1, T2, T1c, and FLAIR sequences, yet sequence data is frequently missing due to heterogeneity in multi-institutional datasets, including variations in acquisition parameters, scanner protocol, time constraints, and image artifacts such as patient motion³⁰⁻³². Under these realistic constraints, maximizing the diagnostic information value of available sequences to achieve performance comparable to complete sequences has become a critical challenge in imaging medicine.

Here, we present **MRICombo**, a unified deep learning-based framework that enables simultaneous anatomical segmentation, tumor localization, grading, and malignancy classification from arbitrary available MRI sequence combinations. A fundamental insight driving our approach is that different MRI sequences and disease types share common underlying imaging features and structural representations. This permits the learning of generalized image representations across multiple disease types. To exploit this property, we introduce a novel mixture-of-experts (MoE) architecture that incorporates sequence-specific feature extraction, adaptive feature fusion, anatomy-task prompting, and gating router to enable synergistic multi-task learning. When evaluated on 12 internal and public datasets, MRICombo achieved state-of-the-art performance across multiple metrics. The framework demonstrated strong generalization on 3 external datasets for anatomical and tumor segmentation, as well as on 1 private dataset for malignancy classification, and exhibited effective transfer learning on 1 unseen anatomical segmentation task. Moreover, the framework retained robust accuracy in missing-sequence scenarios. Collectively, these results validate the framework's strong task handling, generalization, and transfer learning capabilities, highlighting its potential for widespread clinical deployment in real-world imaging settings.

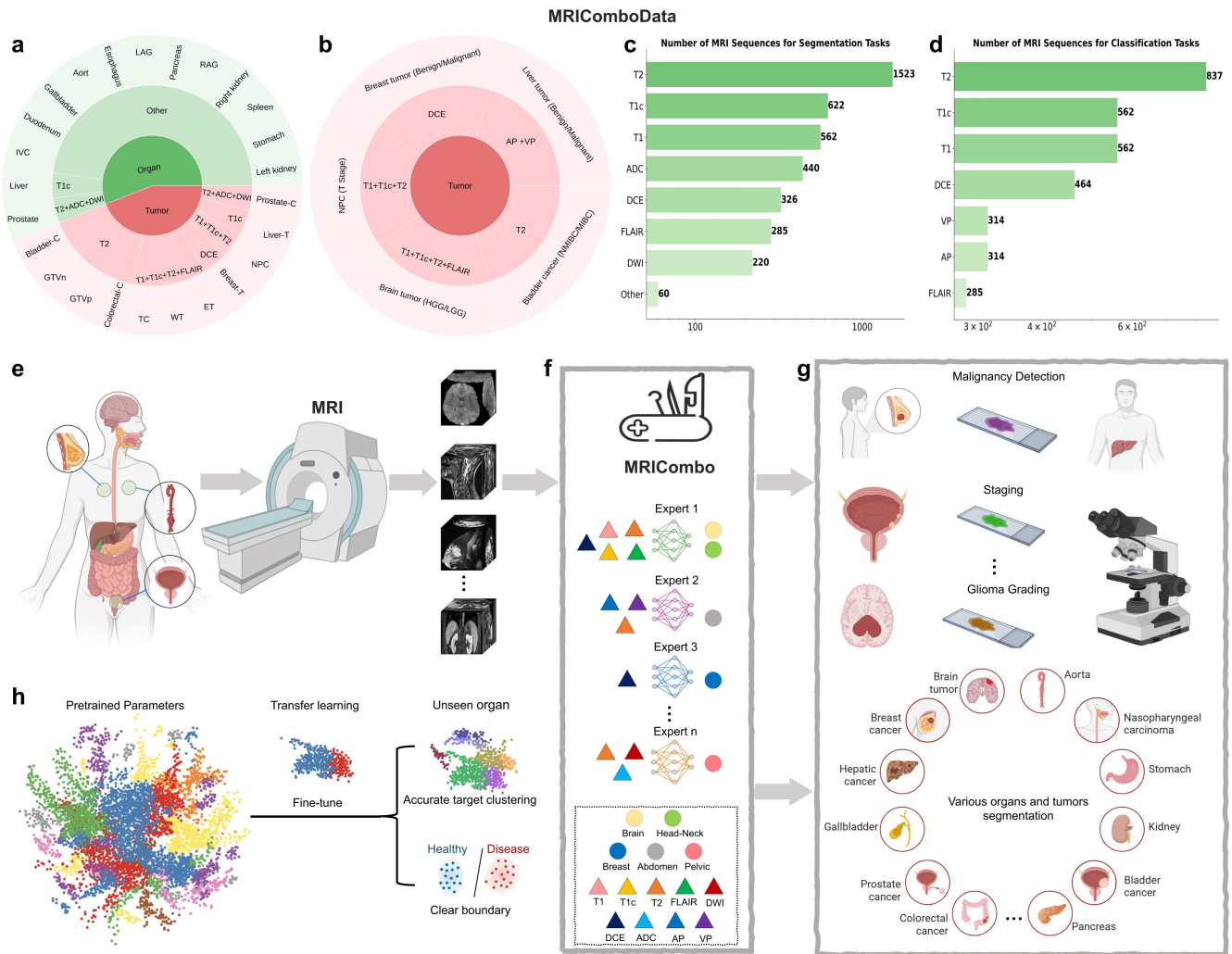


Figure 1. Overview of MRICombo and MRIComboData. **a-d**, Composition and key characteristics of the MRIComboData. **a**, Sunburst chart displaying anatomical regions (organs and tumors) segmented by MRICombo and their corresponding MRI sequences. **b**, Sunburst chart illustrating tumor classification tasks, including benign/malignant discrimination and tumor grading, with associated MRI sequences. **c**, Bar plot quantifying the number of sequences utilized for segmentation tasks. **d**, Bar plot quantifying the number of sequences employed for classification tasks. **e**, The generation of multi-position and multi-sequence MRI data. **f**, Architecture of MRICombo (see Supplementary Fig. 1 for details), featuring multiple expert modules that process diverse anatomical positions and variable MRI sequence combinations. **g**, Outputs of MRICombo across six clinical applications: (i) various organs and tumors segmentation; (ii) breast tumor malignancy detection; (iii) liver tumor malignancy detection; (iv) brain tumor grading (HGG/LGG); (v) bladder cancer invasiveness assessment staging (MIBC/NMIBC); and (vi) nasopharyngeal carcinoma (NPC) T-staging. **h**, Schema of transfer learning based on MRICombo's pretrained parameters for application to unseen organ. Schematics **g** and **h** were created in [BioRender](#).

Abbreviations: LAG, left adrenal gland; RAG, right adrenal gland; IVC, inferior vena cava; WT, whole brain tumor; TC, brain tumor core; ET, enhanced tumor; Bladder-C, bladder cancer; GTVp, primary gross tumor volumes; GTVn, involved metastatic lymph nodes; Colorectal-C, colorectal cancer; Breast-T, breast tumor; Liver-T, liver tumor; AP, arterial phase; VP, venous phase; HGG, high-grade glioma; LGG, low-grade glioma; NMIBC, non-muscle-invasive bladder cancer; MIBC, muscle-invasive bladder cancer.

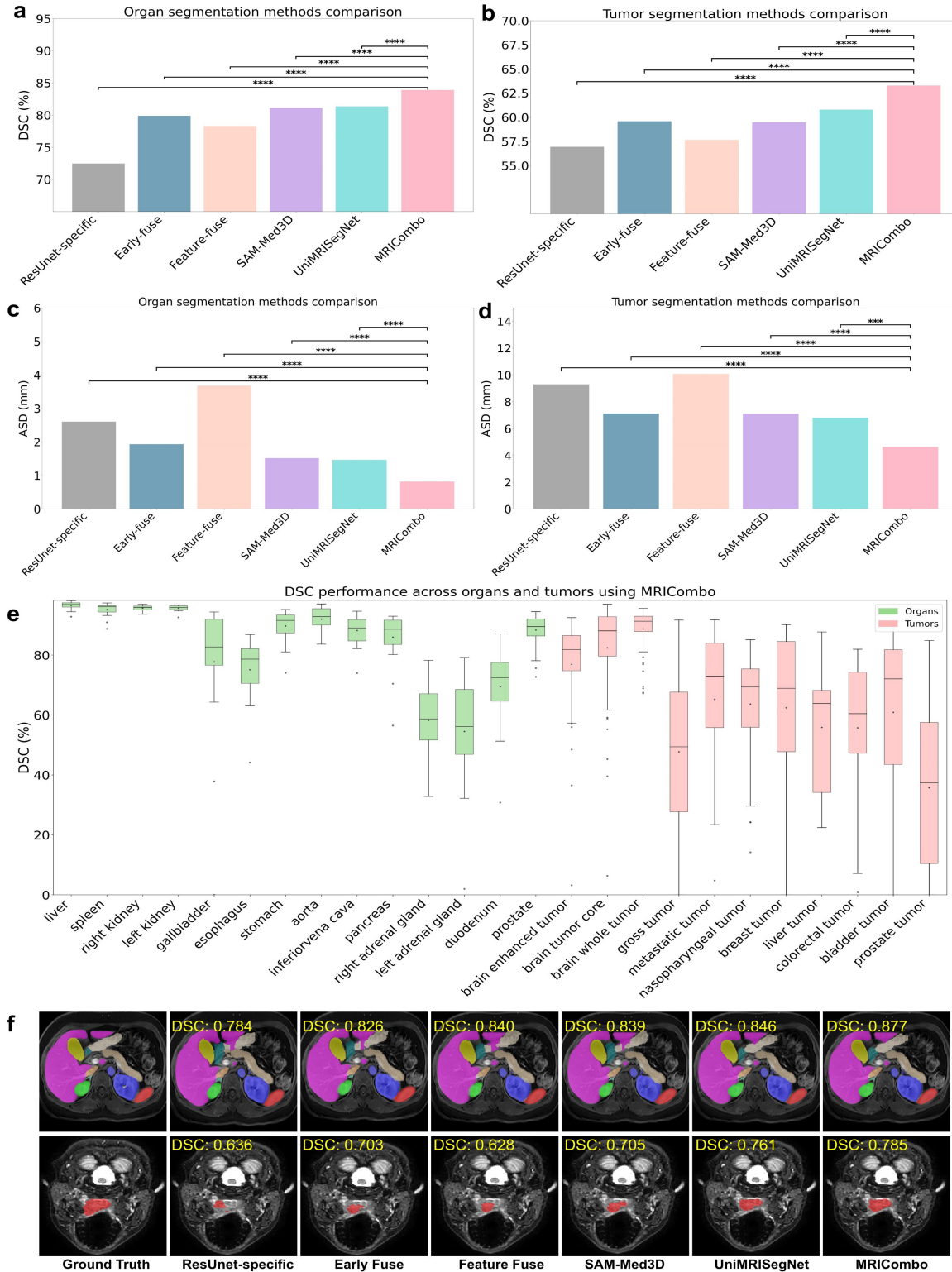


Figure 2. Quantitative and qualitative evaluation of segmentation performance on the internal test set. **a–b**, Bar plots showing Dice Similarity Coefficient (DSC) scores for (a) 14 organ segmentation tasks and (b) 11 tumor segmentation tasks. **c–d**, Bar plots displaying Average Surface Distance (ASD) scores for (c) organ and (d) tumor segmentation tasks. **e**, Comparative analysis of our framework’s average DSC performance across all organ and tumor segmentation tasks. **f**, Representative visualization of segmentation results, showcasing abdominal organs and primary gross tumor volumes (GTVp) across different methods. IPanel (e) presents the results as a box plot, where the lower bound, center line, and upper bound of the box correspond to the first quartile (Q1), median, and third quartile (Q3), respectively, defining the interquartile range (IQR) that contains the central 50% of the data. The whiskers extend to the furthest data points within 1.5×IQR, and any points beyond are classified as outliers. Statistical significance for MRICombo’s superiority over the best-competing method was defined as *** $P < 1 \times 10^{-3}$ and **** $P < 1 \times 10^{-4}$ (two-sided paired t-test).

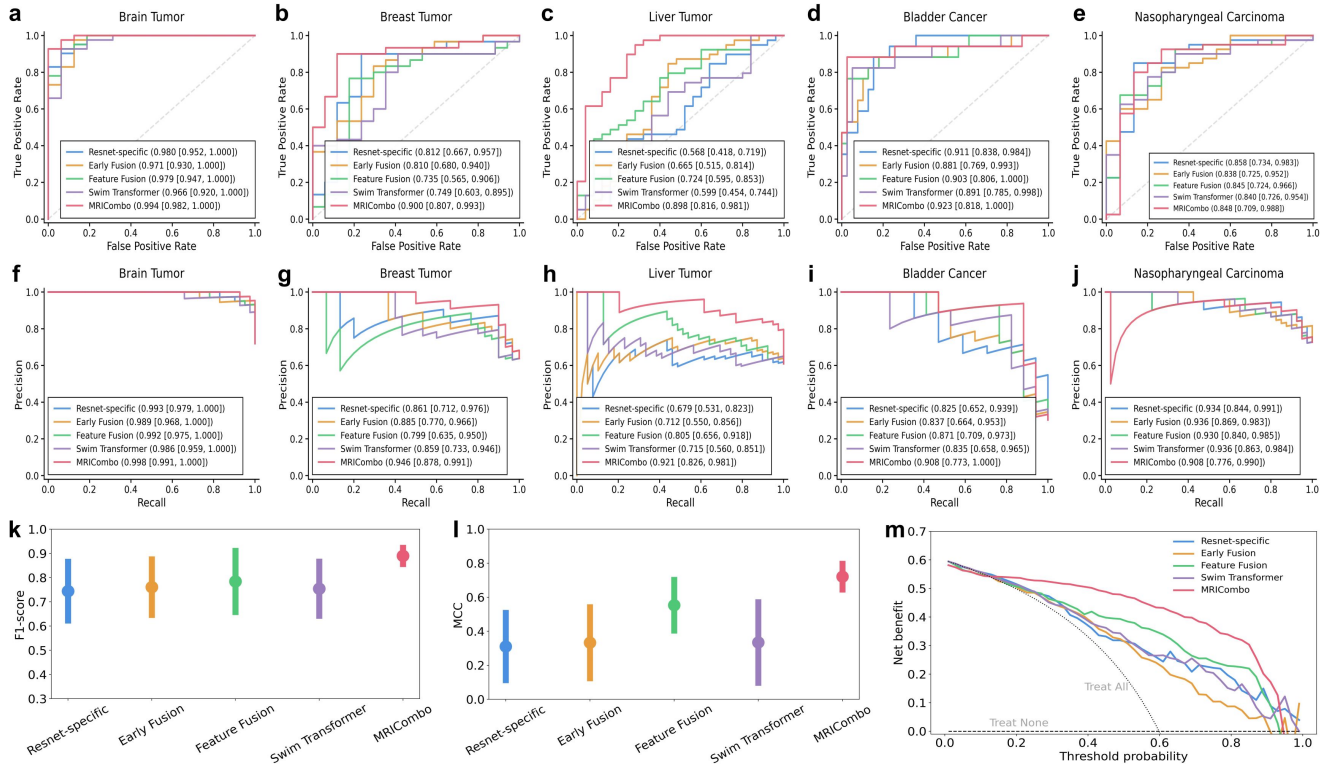


Figure 3. Performance evaluation of MRICombo for malignancy detection and tumor grading on the internal test set. MRICombo achieved competitive AUROC (a–e) and AUPRC (f–j) performance compared to four state-of-the-art methods, outperforming them in four of five tumor types. It also exhibited superior performance in (k) MCC and (l) F1-score. (m) Decision curve analysis further confirmed the enhanced clinical utility of MRICombo across a wide threshold range. All performance curves represent 95% confidence intervals derived from 1,000 bootstrap replicates. MCC, Matthews correlation coefficient; NMIBC, non-muscle-invasive bladder cancer; MIBC, muscle-invasive bladder cancer; NPC, nasopharyngeal carcinoma; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve.

Results

Overview of MRICombo and MRIComboData

MRICombo (Fig. 1f) is designed to serve as a unified framework for universal 3D MRI segmentation, tumor grading, and malignancy classification under varying imaging sequences. A key challenge in building such a framework lies in its ability to adapt to variability in imaging protocols, anatomical contexts, and disease presentations. To address this challenge, we constructed a diverse MRI dataset (MRIComboData) comprising 7,376 scans from 2,282 patients, covering 9 MRI sequences and 25 anatomical structures and tumor types (Fig. 1a–d). Fig. 1a presents a comprehensive taxonomy of segmentation targets and their corresponding MRI sequences used in this study. The anatomical structures and pathological regions are systematically organized into two main categories: Organs (14 types) and Tumors (11 types). The supported MRI sequences include T1, T1c, T2, FLAIR, Dynamic Contrast-Enhanced (DCE; both pre- and post-contrast), and Diffusion-Weighted Imaging (DWI). Single or multi-sequence protocols are driven by the specific clinical task and target anatomy. Fig. 1(b) displays the classification tasks and their corresponding imaging protocols, covering five clinically significant scenarios: (1) brain tumor grading (High-Grade Glioma/Low-Grade Glioma, HGG/LGG), (2) bladder cancer staging (Non-Muscle-Invasive Bladder Cancer/Muscle-Invasive Bladder Cancer, NIBC/MIBC), (3) liver tumor characterization (benign/malignant), (4) breast tumor characterization (benign/malignant), and (5) nasopharyngeal carcinoma T-staging (T1-T2 vs. T3-T4). The associated MRI sequences include T1, T1c, T2, FLAIR, (DCE; both pre- and post-contrast), Arterial Phase (AP), and Venous Phase (VP). These sequences are employed either individually or in combination according to the specific diagnostic requirements of each clinical scenario. Fig. 1c presents the distribution of MRI sequences utilized in segmentation tasks. T2 imaging emerges as the most frequently used sequence with 1,523 instances, followed by T1c (622 cases) and T1 (562 cases). Additional sequences including ADC, DCE, FLAIR, and DWI collectively constitute the total of 4,038 sequences across all modalities. Fig. 1d illustrates the sequence distribution for classification tasks, where T2 imaging remains the predominant choice at 837 instances, followed

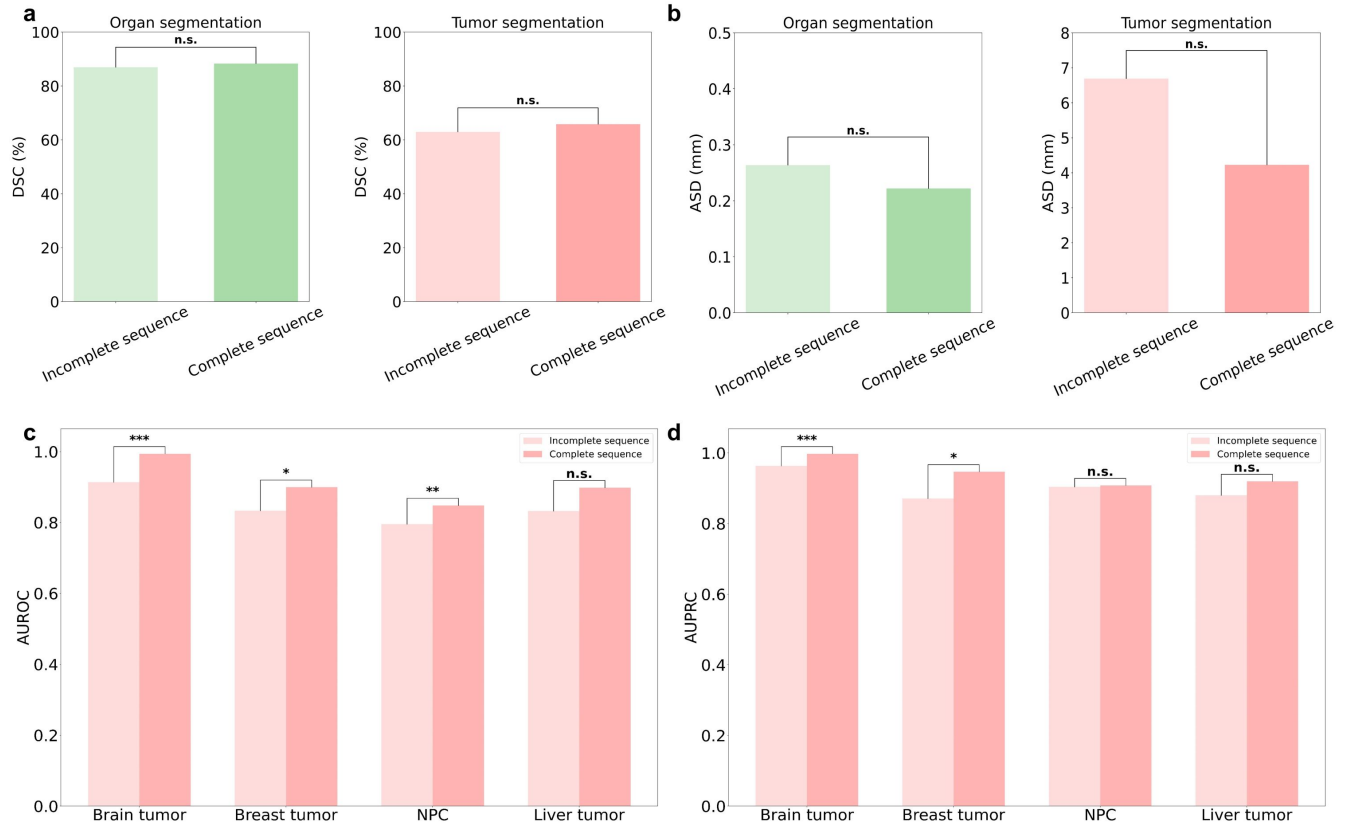


Figure 4. Robustness assessment of MRICombo with missing sequences. Assesses the model’s performance on the internal multi-sequence test set under conditions of input incompleteness. Superiority over the best-competing method was statistically evaluated (two-sided paired t-test; *** $P < 1 \times 10^{-3}$, ** $P < 1 \times 10^{-2}$, n.s.: not significant). Abbreviations: NPC, nasopharyngeal carcinoma.

by T1c and T1 (562 instances each), DCE (464 instances), and dynamic phase imaging encompassing both AP and VP (314 instances each), resulting in a total of 3,338 sequences analyzed. The scale and diversity of the dataset empower MRICombo to learn rich medical imaging representations capable of capturing a wide range of anatomical structures and pathological findings across different sequence modalities. The variability in imaging protocols and corresponding segmentation and classification targets underscores the need for a unified framework that can handle both the distinctive characteristics of individual modalities and their combined features, while maintaining robust performance across diverse clinical tasks. The representations learned from this MRIComboData can be effectively fine-tuned to facilitate transfer learning to unseen organ clusters and disease differentiations (Fig. 1h), demonstrating strong generalization capability.

Superior segmentation performance over task-specific and universal models

A comprehensive comparison was conducted to evaluate our proposed MRICombo framework against five state-of-the-art segmentation approaches: ResUNet-specific³³, Early Fuse, Feature Fuse³⁴, SAM-Med3D²⁰, and UniMRISegNet³⁵ on an internal test set. The segmentation performance was quantified using the Dice Similarity Coefficient (DSC) and the Average Surface Distance (ASD). Our results confirm the statistically significant superiority of the proposed MRICombo framework.

Overall, MRICombo consistently surpassed all competing methods (Supplementary Table 2), achieving average Dice Similarity Coefficient (DSC) scores of 83.90% for organ segmentation and 63.30% for tumor segmentation. On the organ segmentation task, MRICombo outperformed the next-best model, UniMRISegNet, by 2.54% in DSC (Fig. 2a, $p < 0.0001$). For the more challenging tumor segmentation, our framework also secured a superior performance, leading UniMRISegNet by 2.52% in DSC (Fig. 2b, $p < 0.0001$). Furthermore, MRICombo demonstrated exceptional capability in surface detection, attaining average symmetric surface distances (ASD) of 0.65 mm for organs and 4.64 mm for tumors. This represents a significant reduction in ASD of 1.82 mm and 2.17 mm compared to the best existing method, as shown in Fig. 2c ($p < 0.0001$) and Fig. 2d ($p < 0.001$), respectively. These results collectively indicate that MRICombo provides a more powerful segmentation encoder than all benchmarked models.

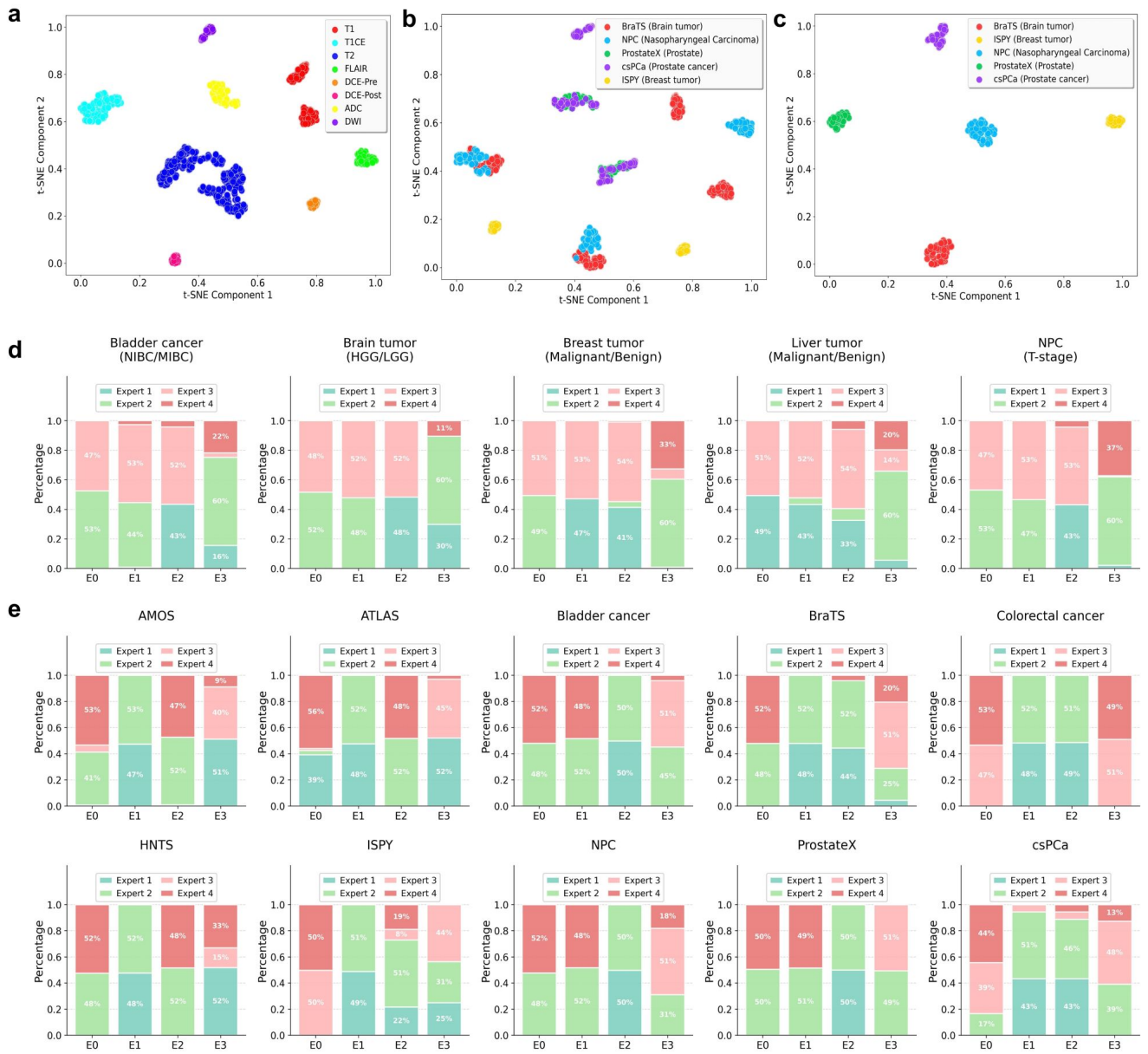


Figure 5. Visualization of feature representations and expert activation patterns. (a) t-SNE visualization of sequence-specific feature representations across all segmentation training data. (b) t-SNE visualization of multi-sequence features from different anatomical regions across five datasets. (c) t-SNE visualization of the adaptive multi-sequence feature fusion module on five datasets. (d) Expert weight distribution across encoder layers (Encoder₀–Encoder₃) for 15 tasks (5 classification and 10 segmentation). (e) Expert weight distribution across encoder layers (Encoder₀–Encoder₃) for 10 segmentation tasks. All analyses employ the Top-2 expert activation strategy, with color coding indicating normalized weights of the four experts. Each subplot corresponds to an individual task, where the horizontal axis represents network layers and the vertical axis shows expert weight proportions. E0–E3 denote Encoder₀–Encoder₃.

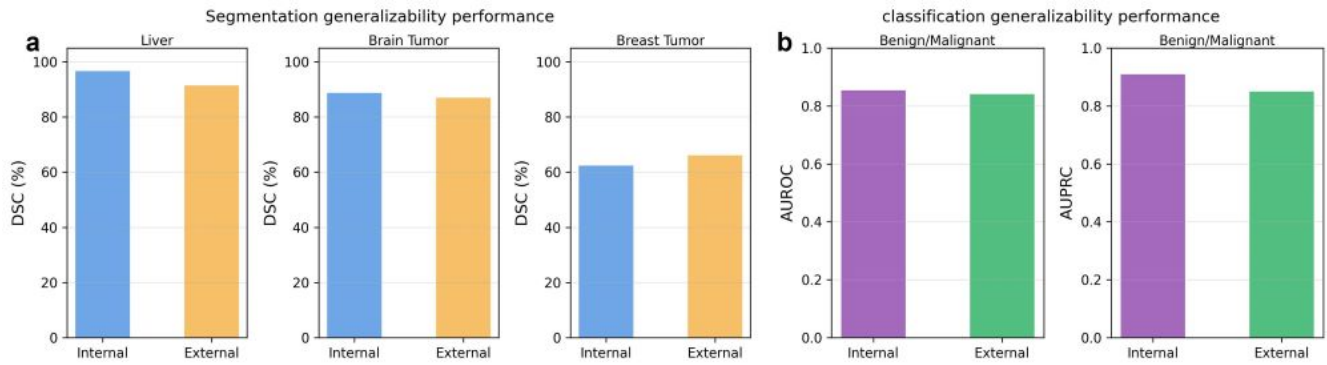


Figure 6. Generalization performance comparison on internal and external test sets. **(a) Segmentation generalizability performance** across three target types (Liver, Brain Tumor, and Breast Tumor) measured by Dice Similarity Coefficient (DSC). The model demonstrates robust performance with minimal degradation when applied to external datasets, showing DSC scores of 96.58% vs 91.42% for liver, 88.69% vs 86.95% for brain tumor, and 62.42% vs 66.05% for breast tumor (internal vs external, respectively). **(b) Classification generalizability performance** for liver tumor benign/malignant classification tasks measured by AUROC and AUPRC metrics. Both metrics show consistent performance between internal and external datasets (AUROC: 0.854 vs 0.841; AUPRC: 0.909 vs 0.849), indicating good model generalization capabilities across different clinical settings.

We further perform comprehensive anatomical structure performance analysis to validate the clinical utility for specific organs and tumors. Detailed analysis across 25 anatomical structures revealed MRICombo's robust performance across diverse organs and tumors (Fig. 2e). MRICombo consistently achieved high DSC values above 95% for four organs: liver, spleen, right kidney, and left kidney. For brain tumors, MRICombo attained the highest DSC values, ranging from 80% to 90%. For tumors in anatomically challenging regions, including pancreatic and colorectal tumors, the framework exhibited lower yet still competitive performance, with DSC values between 40% and 60%.

In addition to quantitative evaluation, qualitative assessment was performed to validate segmentation quality across diverse anatomical locations. Fig. 2f presents representative examples from two challenging scenarios: an abdominal multi-organ case from the AMOS dataset and a nasopharyngeal carcinoma (NPC) case from the NPC dataset. Visual inspection revealed that MRICombo consistently produced segmentations with high fidelity to ground truth annotations across all evaluated regions, whereas competing methods exhibited systematic errors including boundary over-segmentation and volumetric under-segmentation. Notably, alternative approaches failed to accurately delineate "the tumor margins in NPC cases", likely attributable to insufficient modeling of complex spatial relationships and tissue heterogeneity in the tumor microenvironment. SAM-Med3D, despite leveraging foundation model capabilities, exhibited inconsistent performance on organ boundaries with ambiguous intensity transitions. UniMRISegNet showed improved results compared to single-task methods but still struggled with precise tumor margin delineation. In contrast, MRICombo maintained robust segmentation accuracy across both abdominal organs and head-neck tumors, demonstrating superior generalization. These qualitative findings corroborate the quantitative metrics, confirming MRICombo's advantages for anatomically complex and clinically challenging segmentation tasks.

Consistent disease detection performance across five anatomic regions

We systematically evaluated the classification performance of MRICombo against four competing approaches—ResNet-specific³⁶, Early Fusion, Feature Fusion³⁴, and Swin Transformer³⁷—across five distinct tumor classification tasks: brain tumor subtyping (high-grade glioma vs. low-grade glioma), breast tumor characterization (benign vs. malignant), liver tumor differentiation (benign vs. malignant), bladder cancer staging (muscle-invasive vs. non-muscle-invasive), and nasopharyngeal carcinoma T-staging (T1-T2 vs. T3-T4). MRICombo demonstrated consistently superior performance across all tumor types, achieving the highest area under the receiver operating characteristic curve (AUROC) in every classification task (Fig. 3a-e). Specifically, MRICombo attained AUROC values of 0.994 (95% CI: 0.982–1.000) for brain tumors, 0.900 (95% CI: 0.807–0.993) for breast tumors, 0.898 (95% CI: 0.816–0.981) for liver tumors, 0.923 (95% CI: 0.818–1.000) for bladder cancer, and 0.848 (95% CI: 0.709–0.988) for nasopharyngeal carcinoma. These results represent improvements of 1.4%, 8.8%, 17.4%, 1.2%, and modest gains over the second-best performing methods, respectively. Complementary precision-recall analysis corroborated MRICombo's superior diagnostic capability, yielding area under the precision-recall curve (AUPRC) values of 0.998 (95% CI: 0.991–1.000) for brain tumors, 0.946 (95% CI: 0.878–0.991) for breast tumors, 0.921 (95% CI: 0.826–0.981) for liver tumors, 0.908 (95% CI: 0.773–1.000) for bladder cancer, and 0.908 (95% CI: 0.776–0.990) for nasopharyngeal carcinoma (Fig. 3f-j). These findings establish MRICombo's robust generalization capability across diverse tumor types and

anatomical sites.

Comprehensive statistical validation using Matthews Correlation Coefficient (MCC) and F1-score metrics further confirmed MRICombo's performance advantage (Fig. 3k-l). Our framework achieved the highest MCC (0.72 ± 0.09) and F1-score (0.87 ± 0.04) (mean $\pm$ s.d.), indicating stable performance. Feature Fusion and Swin Transformer demonstrated intermediate performance, while ResNet-specific and Early Fusion approaches yielded lower metrics across all evaluations. Decision curve analysis revealed the clinical utility of MRICombo across threshold probabilities ranging from 0.1 to 0.8 (Fig. 3m), demonstrating superior net benefit compared to treat-all or treat-none strategies throughout this range.

Performance under incomplete sequence conditions

Incomplete sequence conditions closely mirror real clinical scenarios where certain MRI sequences may be unavailable due to scan time constraints, patient intolerance, or equipment limitations^{31,32}. We systematically evaluated MRICombo's performance under all possible incomplete sequence combinations using the multi-sequence internal test set (Supplementary Table 1).

Incomplete sequence generation strategy. For datasets with n available sequences, we tested all $2^n - 2$ possible incomplete combinations, excluding only the complete sequence (all present) and null sequence (all absent) scenarios. For example, brain tumor datasets with four sequences (T1, T2, T1-contrast, FLAIR) yielded $2^4 - 2 = 14$ incomplete combinations, including single-sequence scenarios (e.g., T1 only), dual-sequence scenarios (e.g., T1+T2), and triple-sequence scenarios (e.g., T1+T2+FLAIR). This exhaustive evaluation strategy ensures comprehensive assessment of MRICombo robustness across all clinically relevant sequence availability patterns. (See Supplementary Table. 1 for incomplete combinations.)

Segmentation performance. MRICombo demonstrated robust segmentation capabilities across all incomplete sequence combinations (Fig. 4a-b). Averaged across all incomplete sequence combinations, organ segmentation achieved a mean DSC of 86.99% compared to 88.36% with complete sequences, representing only a 1.55% relative performance decrease. Tumor segmentation showed a mean DSC of 63.00% with incomplete sequences versus 65.87% with complete sequences, indicating a 4.36% relative reduction. Overall, the segmentation performance degradation remained below 5% for both organ and tumor segmentation tasks, even in the most challenging single-sequence scenarios.

Surface distance measurements corroborated these findings. MRICombo achieved an average surface distance (ASD) of 0.26 mm across incomplete sequence combinations compared to 0.22 mm with complete sequences for organ segmentation (Fig. 4b). Tumor segmentation exhibited an ASD of 6.69 mm for incomplete sequences versus 4.22 mm for complete sequences. These modest increases demonstrate the model's resilience to missing sequence information.

Classification performance. The classification performance of MRICombo remained competitive across incomplete sequence combinations, with no statistically significant differences observed between incomplete and complete sequences for most tumor types (Fig. 4c-d). Averaged across all incomplete sequence combinations, AUROC measurements for complete versus incomplete sequences were as follows: brain tumor (0.994 vs. 0.914), breast tumor (0.900 vs. 0.833), nasopharyngeal carcinoma (0.848 vs. 0.796), and liver tumor (0.898 vs. 0.832) (Fig. 4c). Corresponding AUPRC values showed similar trends: brain tumor (0.998 vs. 0.963), breast tumor (0.946 vs. 0.871), nasopharyngeal carcinoma (0.908 vs. 0.904), and liver tumor (0.921 vs. 0.880) (Fig. 4d).

The average AUROC and AUPRC reductions across all tumor types were 7.84% and 4.31%, respectively, when averaged over all incomplete sequence combinations. Nasopharyngeal carcinoma classification demonstrated the smallest performance gap (6.13% AUROC reduction), while breast tumor classification exhibited the largest decrease (7.44% AUROC reduction). Brain tumor and liver tumor classifications showed intermediate reductions of 8.04% and 7.35%, respectively. Performance retention analysis revealed differential sequence dependencies across anatomical regions. When averaged across incomplete sequence combinations, brain tumor classification retained 92.0% of complete sequence AUROC performance, breast tumor retained 92.6%, nasopharyngeal carcinoma preserved 93.9%, and liver tumor sustained 92.6% of original performance levels. These results demonstrate MRICombo's resilience to incomplete input data across diverse sequence availability patterns, a critical advantage for real-world clinical deployment.

Interpretability analysis

Fig. 5a presents the t-SNE visualization of the sequence feature extraction module. It can be observed that the sequence feature extraction experts map different anatomical regions with the same sequence type into a unified feature space. Notably, distinct clustering effects are evident for all sequence types except T1, which did not achieve satisfactory clustering performance. Fig. 5b further reveals the inherent heterogeneity of multi-sequence datasets (BraTS, NPC, ProstateX, csPCa, ISPY), showing significant feature space discrepancies between different sequences from the same organ. Fig. 5c illustrates how our fusion module effectively integrates information from heterogeneous sequences into a common feature space. This integration enables the model to focus more effectively on semantic distinctions within the images. After processing by the fusion module, inter-sequence discrepancies are substantially reduced, and features primarily cluster according to their organ origins rather than acquisition protocols..

As shown in Fig. 5(d,e), the expert activation patterns of our framework reveal a consistent routing strategy characterized by shared components and specialized adaptations. This adaptation behaviors are orchestrated by the dual-prompt mechanism of our framework, where the dataset prompt and task prompt collectively guide the expert selection process. A detailed comparison between classification and segmentation tasks demonstrates a *shared-yet-specialized* routing behavior across encoder layers. In the NPC dataset, for example, both tasks activate Expert 2 in Encoder₀–Encoder₁, while their respective task prompts drive specialization: classification engages Expert 3 whereas segmentation selects Expert 4. This pattern persists through deeper layers, with Encoder₂ sharing Expert 1 but differentiating through complementary experts (Expert 3 for classification versus Expert 2 for segmentation), and Encoder₃ sharing Expert 2 while specializing with Expert 4 (classification) versus Expert 3 (segmentation). This structured scheme, where dataset prompts establish common feature extraction pathways while task prompts enable specialized processing, is consistently observed across brain tumor and bladder cancer datasets. This layered routing strategy demonstrates that our framework does not simply assign disjoint sets of experts to different tasks. Instead, at each hierarchical level, through the coordinated action of dataset and task prompts, it preserves a common expert for task-agnostic feature extraction while dynamically recruiting specialized experts to fulfill task-specific requirements. Such a design elegantly balances specialization with efficient cross-task representation sharing.

Beyond task-level commonality, anatomically adjacent organ regions also exhibit strong consistency in expert selection, reflecting *biologically-grounded coherence*. As shown in Fig. 5e, abdominal datasets including AMOS and ATLAS display highly aligned expert activation patterns. This trend is biologically intuitive since organs within the abdominal cavity such as the liver, kidneys and spleen share tissue properties, developmental origins, and imaging characteristics, which encourages the MoE router to adopt common feature extraction strategies. A similar tendency is observed in pelvic regions including bladder cancer, ProstateX and csPCa, as well as head and neck regions including NPC, HNTS and BraTS. These findings indicate that the MoE framework not only learns from task-level supervision but also implicitly captures intrinsic anatomical and imaging relationships, thereby autonomously organizing its representations in a biologically meaningful manner.

Generalization performance on external dataset

Our MRICombo framework demonstrates robust generalization capabilities across both segmentation and classification tasks. As shown in Fig. 6a, the segmentation performance evaluation across three organ types reveals minimal performance degradation when transferring from internal to external datasets. Specifically, liver segmentation achieves a DSC of 96.58% on internal data and 91.42% on external data, maintaining a 94.7% performance retention rate. Brain tumor segmentation shows excellent stability with scores of 88.69% vs 86.95% (98.0% retention), while breast tumor segmentation demonstrates positive transfer with an improvement from 62.42% to 66.05%. These findings confirm that the framework maintains stable segmentation accuracy across different anatomical structures and imaging protocols in external validation datasets.

For the liver tumor benign/malignant classification task, our framework exhibits similar performance as illustrated in Fig. 6b. The AUROC metric shows only a minor decrease from 0.854 on internal data to 0.841 on external data (1.5% degradation), while AUPRC decreases from 0.909 to 0.849 (6.6% degradation). These results indicate that despite potential differences in imaging parameters, patient demographics, and data acquisition protocols in external datasets, our framework maintains high discriminative capability for tumor characterization.

The consistent performance across both tasks demonstrates the effectiveness of our multi-task learning framework in learning generalizable feature representations. The cross-domain stability in segmentation tasks confirms that the learned features possess good transferability across different clinical settings, while the robust classification performance validates the framework’s practical utility for real-world deployment. These findings support the clinical applicability and reliability of our proposed unified framework across diverse medical imaging scenarios.

Table 1. Performance Comparison on Unseen Hippocampus Segmentation Data (DSC, %)

Setting	Model	Anterior hippocampus	Posterior hippocampus	Average
Scratch	ResUnet	87.01	84.81	85.91
Transfer	SAM-Med3D	60.99	56.70	58.85
	UniMRISegNet	87.42	84.84	86.13
	MRICombo (ours)	88.32	85.94	87.13

Evaluation of transfer learning on unseen data domains

Table 1 compares the generalization performance of different models on unseen hippocampus segmentation data under transfer learning settings. While the general-purpose model SAM-Med3D²⁰ achieves only 58.85% in average DSC—significantly

lagging behind the specialized ResUnet trained from scratch (85.91%)—both UniMRISegNet³⁵ and our proposed MRICombo exhibit strong transfer learning capability. Specifically, UniMRISegNet attains an average DSC of 86.13%, slightly surpassing the scratch-trained ResUnet, whereas our MRICombo establishes a new state-of-the-art result of 87.13%. This corresponds to an improvement of 1.22% over the ResUnet baseline and 1.00% over UniMRISegNet.

The superior transfer ability of our model can be largely attributed to its capacity for learning shared representations across different organs within the same MRI sequence. In contrast, SAM-Med3D was primarily trained on a combination of CT and MRI data, which may limit its sensitivity to modality-specific features in MRI and hinder its adaptability to unseen MRI domains.

Table 2. Ablation study of MRICombo. Classification performance is measured by AUROC, while segmentation performance is evaluated using (DSC, %).

Method	Classification (AUROC)					Average	Segmentation (DSC, %)		Average
	Brain	Liver	Breast	Bladder	Nasopharynx		Organs	Tumors	
Classification Baseline	0.971	0.665	0.810	0.881	0.838	0.833	-	-	-
Segmentation Baseline	-	-	-	-	-	-	79.90	59.58	69.74
Multi-task	0.962	0.521	0.785	0.883	0.832	0.797	77.15	56.28	66.72
Multi-task + Adaptive Fusion	0.983	0.755	0.815	0.892	0.841	0.857	80.06	60.02	70.04
Multi-task + MoE	0.986	0.835	0.865	0.908	0.839	0.887	82.24	62.72	72.48
MRICombo	0.994	0.898	0.900	0.923	0.848	0.913	83.90	63.30	73.60

Table 3. Comparison of computational complexity across different models in segmentation task.

Model	GPU Memory Usage (G)	Params (M)	Inference Time (s)
ResUnet-specific	$6.2 \times N$	$40.6 \times N$	0.50
Early fuse	$7.9 \times N$	$41.2 \times N$	0.51
Feature fuse	$9.7 \times N$	$100.6 \times N$	0.64
SAM-Med3D	8.2	100.5	3.32
UniMRISegNet	9.1	14.2	0.58
MRICombo (ours)	15.9	118.2	0.53

Ablation study

We conducted a comprehensive ablation study to systematically evaluate the individual and synergistic contributions of the adaptive fusion and Mixture of Experts (MoE) modules within our proposed framework. All experiments were performed using a universal single-task architecture as the baseline to ensure fair comparison. As shown in Table 2, our proposed modules demonstrate substantial performance improvements over this baseline.

The universal single-task baselines established competitive performance benchmarks: the classification baseline achieved an average area under the receiver operating characteristic curve (AUROC) of 0.833, while the segmentation baseline attained an average Dice similarity coefficient (DSC) of 69.74%. Joint training with a shared backbone network, however, led to performance degradation across both tasks. The average classification AUROC decreased to 0.797 (a 3.6% relative reduction), while the average segmentation DSC dropped to 66.72%, with tumor segmentation declining to 56.28% DSC (a 3.0% relative reduction from the single-task baseline). This deterioration stems from gradient conflicts between these heterogeneous tasks. Segmentation, particularly for precise tumor boundary delineation, demands localized, fine-grained spatial features, whereas classification relies on global, context-aware semantic representations. When constrained to share a common encoder, gradients from competing objectives propagate in divergent directions, forcing a suboptimal compromise.

The introduction of adaptive fusion alone showed moderate recovery in classification performance (AUROC: 0.857). However, tumor segmentation remained suboptimal (60.02% DSC) compared with the single-task baseline, indicating that adaptive fusion alone cannot fully resolve the fundamental feature representation conflicts between tasks. The incorporation of the MoE module effectively addressed this limitation. The Multi-task + MoE configuration not only recovered performance losses but substantially exceeded all baselines, achieving average AUROC of 0.887 and average DSC of 72.48%. These results demonstrate that MoE enables specialized, task-adapted feature processing through dedicated expert pathways, effectively alleviating task interference through expert-guided specialization. The complete framework, which sequentially integrates adaptive fusion followed by MoE processing, achieved the highest performance with an average classification AUROC of 0.913

and an average segmentation DSC of 73.60%. These results demonstrate a clear synergistic interaction: adaptive fusion first effectively integrates multi-scale features, upon which the MoE module subsequently applies specialized expert transformations to generate disentangled, task-specific representations. This sequential processing pipeline enables robust performance across diverse anatomical regions and task types.

In summary, the proposed modules deliver substantial improvements over the universal baseline. While adaptive fusion provides initial gains, the MoE mechanism is crucial for preventing fundamental task conflicts through expert specialization. Their integration yields a robust universal model capable of excelling at both classification and segmentation tasks.

Lower computational complexity compared to task-dependent models

As shown in Table 3, we compare the computational efficiency of different models on the segmentation task, categorizing them into task-dependent models (top section) and unified models (bottom section). The computational resource requirements of task-dependent models scale linearly with the number of tasks N , whereas unified models maintain fixed resource usage independent of N .

A specific comparison with the ResUnet-specific model underscores the advantages of our proposed MRICombo in multi-task deployment scenarios. While a single ResUnet-specific model requires 6.2 G of GPU memory and 0.50 s for inference, deploying multiple such task-specific models results in linearly growing resource demands—specifically, $6.2 \times N$ G memory and $40.6 \times N$ M parameters. In contrast, MRICombo maintains a constant computational footprint of 15.9 G GPU memory and 118.2 M parameters, regardless of the number of tasks, achieving substantial resource savings when $N \geq 3$ (in this study, $N = 15$).

Notably, MRICombo’s inference time of 0.53 s shows only a minimal 6% increase compared to the ResUnet-specific model (0.50 s), while providing comprehensive multi-task capability. This marginal time overhead is well justified by eliminating the need to maintain and execute multiple separate models, thereby significantly simplifying the deployment architecture and reducing system complexity. Moreover, such a low-footprint model is essential for enabling direct integration and real-time execution on MRI scanners, where hardware resources are often constrained and computational headroom is limited.

Discussion

We introduce MRICombo, a unified multi-task deep learning framework for heterogeneous clinical MRI analysis. It performs universal anatomical segmentation, tumor localization, grading, staging, and malignancy detection across multiple organ systems within a single architecture. Unlike conventional task-specific or universal models, MRICombo employs a dual-prompt guided mixture-of-expert (MoE) architecture where dataset prompts and task prompts dynamically modulate expert routing based on task types and clinical objectives. This design stems from the insight that diverse MRI sequences and disease manifestations share underlying anatomical and imaging patterns, which enables the learning of highly transferable representations across diseases and body regions. Our results show that MRICombo achieves state-of-the-art performance across diverse tasks (Fig. 2, Fig. 3, Fig. 4, and Fig. 6), while exhibiting notable robustness to incomplete imaging sequences, a critical requirement for real-world clinical deployment.

A distinguishing feature of MRICombo is its resilience to incomplete sequence availability. Clinical MRI protocols are inherently heterogeneous due to scan time constraints, patient intolerance, equipment variations, and institutional preferences^{30–32}. MRICombo maintained high diagnostic accuracy with minimal performance decline, showing an average AUROC reduction below 8% and an average DSC reduction below 5% under missing sequence conditions. This robustness is clinically significant, as it enables the framework to accommodate existing clinical workflows without requiring protocol standardization, a major barrier to artificial intelligence (AI) adoption in routine practice.

Architecturally, MRICombo offers fundamental insights into multi-task learning principles in medical imaging. Our analysis reveals a sophisticated *shared-yet-specialized* routing mechanism that operates hierarchically across network layers. Rather than employing disjoint expert sets for different tasks, the framework maintains a core of shared experts for fundamental feature extraction while dynamically engaging specialized experts to address task-specific requirements. This balanced approach enables efficient knowledge sharing across tasks while preserving the unique characteristics needed for different clinical objectives, resulting in a more robust and scalable architecture for diverse clinical tasks. Beyond task-level optimization, MRICombo demonstrates remarkable *biologically-grounded coherence* in its learned representations. The framework spontaneously organizes its expert selection patterns according to anatomical relationships, with adjacent organ regions consistently activating similar expert combinations. This emergent property indicates that the framework successfully captures the intrinsic biological continuity of human anatomy, developing feature representation strategies that respect real-world anatomical organization. Such biologically meaningful organization not only enhances framework interpretability but also suggests that the framework learns representations aligned with fundamental anatomical principles. Furthermore, MRICombo effectively processes multi-sequence data through a dedicated architecture. Specialized sequence module extract unified features within each sequence type, while an adaptive fusion module synthesizes these heterogeneous inputs into a semantically coherent representation. This process

emphasizes diagnostically relevant information, enabling the framework to operate robustly across diverse imaging protocols and institutional practices.

From a translational perspective, MRICombo offers practical advantages for clinical integration. Computational analysis demonstrated that for applications exceeding three tasks, encompassing 15 tasks in this study, the unified framework substantially reduces storage and computational requirements compared to deploying multiple task-specific models, while maintaining comparable inference speeds. This efficiency is particularly relevant for deployment scenarios including embedded scanner implementations for real-time analysis, cloud-based teleradiology platforms, and resource-limited healthcare settings. By consolidating multiple diagnostic functions, the framework may reduce repetitive workflow burdens on radiologists, enabling greater focus on complex clinical reasoning and patient management.

Several limitations warrant consideration. First, the dataset's sequence imbalance, with a predominance of T2-weighted imaging, may hinder performance on tasks requiring underrepresented sequences like DWI for prostate or multi-phase liver assessment. Second, segmentation of small lesions (e.g., in prostate cancer or bladder cancer) remains challenging due to low contrast and limited volume, an issue common across methods. Finally, the framework currently relies solely on imaging data, which future work should aim to augment with clinical metadata to enhance diagnostic and prognostic utility.

In summary, this work establishes the viability of a unified framework capable of performing multiple segmentation and classification tasks across heterogeneous MRI sequences, eliminating the need for specialized models. The proposed architecture achieves specialist-level performance while accommodating diverse clinical tasks and imaging protocols, signaling a shift toward integrated systems for more precise and efficient oncology care.

Methods

Study population and data acquisition

Training and internal validation cohorts. This study utilized data from 12 publicly available MRI datasets, collectively termed MRIComboData, comprising 2,282 participants with 7,376 total scans for model training and internal validation. All datasets were obtained from established repositories including Zenodo, The Cancer Imaging Archive (TCIA)³⁸, Medical Image Computing and Computer Assisted Intervention Society (MICCAI) challenges, GitHub, and Kaggle. Ethical approval and informed consent for the original data collection were obtained by the respective contributing institutions, as documented in their primary publications. This retrospective analysis of de-identified public data was exempt from additional institutional review board approval. The training cohort encompassed diverse anatomical regions and clinical tasks (Fig. 1a–d, Supplementary Table 1). Segmentation tasks spanned five anatomical domains: (1) brain (whole tumor, tumor core, enhancing tumor); (2) head-neck (nasopharyngeal carcinoma, primary gross tumor volumes, involved metastatic lymph nodes); (3) breast (primary tumor); (4) abdomen (liver, spleen, right kidney, left kidney, gallbladder, esophagus, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal glands, duodenum); and (5) pelvis (colorectal cancer, bladder cancer, prostate, prostate cancer). These encompassed 25 distinct segmentation targets across 11 major tumor types and 14 critical organ structures. Classification tasks focused on tumor characterization in brain, breast, liver, bladder, and nasopharyngeal regions. Each participant's MRI data included multiple sequences (range: 1–4 sequences per participant), with sequence types varying by anatomical region and clinical indication (Supplementary Table 1).

External validation cohorts. To evaluate model generalizability across independent datasets, we assembled four external validation cohorts totaling 734 participants (Table 1). These comprised three segmentation datasets and one classification dataset from geographically and technically diverse sources. The external cohorts included: (1) Private-Liver (n = 506), a proprietary dataset of hepatobiliary phase (HBP) MRI scans for liver segmentation, collected under institutional review board approval (Ethics Committee of Nanfang Hospital, Southern Medical University, approval number NFEC-2025-477); (2) BrTS2025-SSA (n = 60), a multi-sequence brain tumor segmentation dataset (T1-weighted, T1-contrast-enhanced, T2-weighted, FLAIR) from sub-Saharan Africa; (3) NACT (n = 64), a breast tumor segmentation dataset featuring dynamic contrast-enhanced (DCE) MRI sequences; and (4) LLD-MMRI2025 (n = 104), a breast tumor classification dataset with arterial phase (AP) and venous phase (VP) imaging. Critically, none of these datasets or their institutional sources were represented in the training cohort, ensuring rigorous evaluation of cross-population generalization.

Unseen anatomy evaluation. To assess transfer learning capability on anatomical structures absent from training data, we employed the Medical Segmentation Decathlon (MSD) Hippocampus dataset (n = 260) as an unseen test set. This dataset comprises T1-weighted MRI scans with expert-annotated hippocampal segmentations. The hippocampus represents a neuroanatomical structure fundamentally distinct from any target in the training cohort, providing a stringent test of the model's ability to generalize learned feature representations to novel anatomical contexts with fine-tuning.

Annotation standards. All segmentation masks and classification labels across training, validation, and test cohorts were derived from expert radiologist annotations. For public datasets, annotations were previously validated through inter-rater reliability assessments or multi-center consensus protocols, as reported in primary publications¹⁸. The Private-Liver dataset annotations were performed by two board-certified radiologists with >3 years of abdominal imaging experience, with

discrepancies resolved by consensus review. All annotations were adopted without modification for model development and evaluation.

Dataset pre-processing

Training deep learning models on diverse and heterogeneous medical imaging datasets presents three primary challenges. First, the framework must simultaneously support end-to-end segmentation and classification within a unified architecture. Second, each task encompasses multiple clinically relevant sub-tasks: segmentation involves delineating various organs and tumors across different anatomical regions, while classification requires distinguishing tumor types and grading categories across multiple sites. Third, the incorporated datasets exhibit substantial heterogeneity in imaging protocols, including variations in sequence types (single-sequence versus multi-sequence studies) and acquisition parameters.

To address these challenges, we implemented a standardized preprocessing pipeline across all datasets. All volumetric images were aligned to a common anatomical coordinate system and resampled to a uniform voxel spacing of $1.2 \times 1.5 \times 1.5$ mm using linear interpolation. Intensity normalization was then performed using z-score normalization, ensuring zero mean and unit standard deviation across each entire volume.

For segmentation tasks, we adopted a patch-based approach, extracting overlapping patches of $96 \times 96 \times 96$ voxels from the preprocessed volumes. This strategy balances computational efficiency with preservation of spatial context for precise boundary delineation. For classification tasks, entire images were resampled to a consistent spatial size, maintaining global anatomical information while ensuring uniform input dimensions across all samples. This dual-path preprocessing framework effectively handles data heterogeneity while enabling simultaneous optimization for both segmentation and classification objectives.

Network architecture

Our framework is built upon a 3D ResUNet³³ architecture with sequence-specific feature extraction, Mixture of Experts (MoE), adaptive multi-sequence feature fusion, and anatomy-task prompt module, as illustrated in Supplementary Fig. 1.

This framework accepts eight MRI sequences as input, including T1-weighted, T1-contrast enhanced (T1c), T2-weighted, FLAIR, diffusion-weighted imaging (DWI), apparent diffusion coefficient (ADC), dynamic contrast-enhanced (DCE), arterial phase (AP), and portal venous phase (VP) sequences. To address the clinical reality where sequence availability varies significantly across patients and institutions, we implement a sequence-aware processing mechanism. Each input sequence (with a fixed length of eight) is accompanied by a binary sequence code vector that indicates sequence availability (1 for available, 0 for missing). This approach eliminates the need for data imputation while maintaining model robustness across different clinical scenarios. The sequence code is processed through a learnable embedding layer with 256 channels to generate sequence-specific feature representations that guide the subsequent stages of the processing pipeline.

Following the standardized processing module, features are extracted through task-specific pathways. We maintain separate feature extraction networks for segmentation and classification tasks, allowing each pathway to develop specialized representations optimized for its respective objective. We employed a multi-sequence feature extraction module to derive image features from different MRI sequences. The module comprises eight separate blocks, each containing a single convolutional layer. For each input volume sequence $X_s \in \mathbb{R}^{H \times W \times D}$, the corresponding convolutional block encodes X_s into a 32-channel feature representation. The resulting features from all sequences are then modulated by their respective sequence codes and subsequently aggregated by summation, producing a 32-channel fused feature map. The fused image feature map combined with two prompts, i.e., task and dataset prompts, are transformed to an gating router, whose weights are used to select Top- k experts.

Both segmentation and classification pathways employ hierarchical MoE structures that enable dynamic adaptation to diverse input patterns. Each MoE layer consists of four expert networks and a gating router that determines expert contributions. Specifically, the feature maps from the preceding MoE layer are first encoded, processed through global average pooling followed by a two-layer multilayer perceptron with ReLU activation³⁹ and a SoftMax layer to generate expert weights. The outputs are then aggregated by summation along the expert channels. For the segmentation pathway, we implement an encoder-decoder structure with MoE components at multiple resolution levels. The encoder employs four experts at each level, where each expert consists of four down-sampling operations using strided convolutions and four residual blocks with instance normalization⁴⁰ and ReLU activation. The decoder mirrors this design with up-sampling operations, integrating skip connections from corresponding encoder levels. The final MoE layer of the decoder aims for expert selection and estimate the probability map $\hat{P} \in \mathbb{R}^{H \times W \times D}$. For the classification pathway, we utilize the encoder with MoE layers without a decoder. Before the classification head, the features from the final layers of the encoders are processed through global average pooling followed by a dropout layer, where we set the dropout rate to 0.2. The input and output sizes of the fully connected layer is 512 and 2, respectively. To enhance computational efficiency while maintaining model expressiveness, we employ the top-2 sparse gating mechanism that only the two most relevant experts are activated for each MoE layer.

448 Compared methods

449 The compared methods encompass both task-specific architectures and general-purpose models. The task-specific benchmarks
 450 include ResUnet-Specific³⁶ for segmentation and Resnet-Specific³³ for classification. Among the general models applicable
 451 to all tasks, we investigate two fusion strategies: Early Fusion, which integrates information by averaging input images at
 452 the network’s input stage, and Feature Fusion, which combines unimodal representations through averaging of high-level
 453 features. Additionally, we evaluate established general segmentation models: SAM-Med3D²⁰ for comprehensive 3D medical
 454 image segmentation and UniMRISegNet³⁵ specifically for MRI data. Finally, the Swin Transformer serves as a flexible
 455 general-purpose backbone architecture. In this work, it is specifically configured for classification tasks, with its outputs adapted
 456 to meet task-specific requirements.

457 Training protocol and experimental setting

458 During training, we extracted patches of voxels from the 3D volume centered around a seed point. A series of randomly chosen
 459 augmentations, such as cropping, rotation ($\pm 15^\circ$), scaling (0.9-1.1), and adjustments to brightness (± 0.2), contrast (± 0.2), and
 460 gamma (0.8-1.2), were applied throughout the training. We use ADAM optimizer as the default optimizer. The initial learning
 461 rate was set to $3e-5$ and the weight decay was $3e-5$. During inference, we employ a sliding window approach with an overlap of
 462 50% between adjacent patches, followed by a weighted averaging of the predictions based on the distance from the patch center
 463 to mitigate boundary artifacts.

464 Loss function

465 Our framework was trained by minimizing the loss function ($\mathcal{L}$) composed of two loss terms: segmentation loss (SL) ($\mathcal{L}_{\text{seg}}$)
 466 and classification loss (CL) ($\mathcal{L}_{\text{cls}}$). For the segmentation task, SL combines binary cross-entropy loss and Dice loss to address
 467 class imbalance issues common in medical image segmentation. Specifically, let $\hat{P}$ and P denote the segmentation prediction
 468 and ground truth, respectively, where $\hat{p}_i$ and p_i represent their corresponding values at voxel i . Therefore, our $\mathcal{L}_{\text{seg}}$ term was:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}} = -\frac{1}{V} \sum_{i=1}^V [p_i \log \hat{p}_i + (1 - p_i) \log(1 - \hat{p}_i)] + 1 - \frac{2 \sum_{i=1}^V p_i \hat{p}_i}{\sum_{i=1}^V (p_i)^2 + \sum_{i=1}^V (\hat{p}_i)^2}, \quad (1)$$

469 where V is the number of voxels in image I . For the classification task, CL is the cross-entropy loss and we defined $\mathcal{L}_{\text{cls}}$ as
 470 follows:

$$\mathcal{L}_{\text{cls}} = - \sum_{j=1}^K y_j \log \hat{y}_j, \quad (2)$$

471 where $\hat{\mathbf{y}} \in \mathbb{R}^K$ represent the predicted probability distribution over K classes, and $\mathbf{y} \in \mathbb{R}^K$ denote the one-hot encoded ground
 472 truth label.

473 Both the two loss terms did not take expert information into account. To encourage uniform expert utilization and prevent
 474 expert collapse, we introduce a load balancing loss for MoE module. Let $\mathbf{g} \in \mathbb{R}^E$ denote the gating weights for E experts. The
 475 load balancing loss uses KL-divergence to measure the deviation from uniform distribution, as follows:

$$\mathcal{L}_{\text{balance}} = \sum_{j=1}^E w_j \log(w_j \cdot E), \quad (3)$$

476 where $w_j = \mathbb{E}[g_j]$ is the average gating probability for expert j across all tokens in a batch. Comining all terms, our overall
 477 loss function ($\mathcal{L}$) was:

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \alpha \mathcal{L}_{\text{cls}} + \beta \mathcal{L}_{\text{balance}}, \quad (4)$$

478 where $\alpha = 0.5$ is the classification weight coefficient that balances the segmentation and classification tasks, and $\beta = 0.05$
 479 is the load balancing coefficient that controls the strength of expert utilization regularization. These hyperparameters were
 480 empirically determined to achieve optimal performance across both tasks. This joint optimization strategy enables the model
 481 to simultaneously learn accurate boundaries of segmented regions, their corresponding class information, and optimal expert
 482 routing, thereby enhancing the overall performance in multi-task scenarios while maintaining computational efficiency through
 483 balanced expert utilization.

Evaluation metrics

Segmentation Metrics. For segmentation evaluation, we followed the recommendations in Metrics Reloaded⁴¹ and employed two complementary metrics: the Dice Similarity Coefficient (DSC) ranging from 0 to 100% and the Average Surface Distance (ASD). These metrics collectively capture both volumetric accuracy and boundary precision. The DSC quantifies the spatial overlap between the predicted segmentation $\hat{P}$ and the ground truth P :

$$\text{DSC}(P, \hat{P}) = \frac{2|P \cap \hat{P}|}{|P| + |\hat{P}|}. \quad (5)$$

The ASD measures the average discrepancy between the boundaries of the prediction and the ground truth:

$$\text{ASD}(P, \hat{P}) = \frac{1}{2} \left[\frac{1}{|\partial P|} \sum_{p \in \partial P} \min_{q \in \partial \hat{P}} \|p - q\| + \frac{1}{|\partial \hat{P}|} \sum_{q \in \partial \hat{P}} \min_{p \in \partial P} \|q - p\| \right]. \quad (6)$$

Classification Metrics. To measure MRICombo's classification performance, the area under the receiver operating characteristic (AUROC), the area under the precision-recall curve (AUPRC), F1-score, and the Matthews's correlation coefficient (MCC) were used to provide comprehensive evaluations. The 95% confidence intervals for the AUC values were calculated using DeLong's method for comparing correlated ROC curves, while confidence intervals for the precision-recall curves were estimated using the non-parametric percentile bootstrap method with 1,000 resamples.

Statistical analysis

All statistical analyses employed two-tailed tests, with p-values of 0.05 or lower deemed significant. These analyses were conducted using Python, version 3.9.7 and scikit-learn version 1.1.3.

Code availability

The complete source code for this study is available at (<https://github.com/zhangzhuoneng/MRICombo>).

References

1. van Beek, E. J. et al. Value of mri in medicine: More than just another test? *J. Magn. Reson. Imaging* **49**, e14–e25 (2019).
2. Radiopaedia. Mri sequences (overview). <https://radiopaedia.org/articles/mri-sequences-overview> (2022).
3. Menze, B. H. et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging* **34**, 1993–2024 (2014).
4. Antonelli, M. et al. The medical segmentation decathlon. *Nat. communications* **13**, 4128 (2022).
5. Zacharaki, E. I. et al. Classification of brain tumor type and grade using mri texture and shape in a machine learning scheme. *Magn. Reson. Medicine: An Off. J. Int. Soc. for Magn. Reson. Medicine* **62**, 1609–1618 (2009).
6. Naser, M. A. & Deen, M. J. Brain tumor segmentation and grading of lower-grade glioma using deep learning in mri images. *Comput. biology medicine* **121**, 103758 (2020).
7. Qiao, M. et al. Breast tumor classification based on mri-us images by disentangling modality features. *IEEE J. Biomed. Heal. Informatics* **26**, 3059–3067 (2022).
8. Gross, M. et al. Automated mri liver segmentation for anatomical segmentation, liver volumetry, and the extraction of radiomics. *Eur. Radiol.* **34**, 5056–5065 (2024).
9. Goel, A. et al. Deployed deep learning kidney segmentation for polycystic kidney disease mri. *Radiol. Artif. Intell.* **4**, e210205 (2022).
10. Zhang, Z. et al. Large-scale multi-center ct and mri segmentation of pancreas with deep learning. *Med. image analysis* **99**, 103382 (2025).
11. Amri, Y., Slama, A. B., Mbarki, Z., Selmi, R. & Trabelsi, H. Automatic glioma segmentation based on efficient u-net model using mri images. *Intell. Medicine* **11**, 100216 (2025).
12. Said, D. et al. Semiautomated segmentation of hepatocellular carcinoma tumors with mri using convolutional neural networks. *Eur. radiology* **33**, 6020–6032 (2023).
13. Yao, L. et al. A colorectal coordinate-driven method for colorectum and colorectal cancer segmentation in conventional ct scans. *IEEE transactions on neural networks learning systems* **36**, 7395–7406 (2024).

14. Chen, Q. et al. Modality-specific information disentanglement from multi-parametric mri for breast tumor segmentation and computer-aided diagnosis. *IEEE Transactions on Med. Imaging* **43**, 1958–1971 (2024).
15. Zhao, J. et al. United adversarial learning for liver tumor segmentation and detection of multi-modality non-contrast mri. *Med. image analysis* **73**, 102154 (2021).
16. Hirsch, L. et al. Radiologist-level performance by using deep learning for segmentation of breast cancers on mri scans. *Radiol. Artif. Intell.* **4**, e200231 (2021).
17. Sumkin, J. H. et al. Diagnostic performance of mri, molecular breast imaging, and contrast-enhanced mammography in women with newly diagnosed breast cancer. *Radiology* **293**, 531–540 (2019).
18. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
19. Cheng, J. et al. Sam-med2d (2023). [2308.16184](https://arxiv.org/abs/2308.16184).
20. Wang, H. et al. Sam-med3d: towards general-purpose segmentation models for volumetric medical images. In *European Conference on Computer Vision*, 51–67 (Springer, 2024).
21. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
22. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. medicine* **30**, 850–862 (2024).
23. Hua, S., Yan, F., Shen, T., Ma, L. & Zhang, X. Pathoduet: Foundation models for pathological slide analysis of h&e and ihc stains. *Med. Image Analysis* **97**, 103289 (2024).
24. Wang, M. et al. Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model. *Nat. Commun.* **16**, 5528 (2025).
25. Ling, Y. et al. Mtanet: Multi-task attention network for automatic medical image segmentation and classification. *IEEE Transactions on Med. Imaging* **43**, 674–685 (2023).
26. Cheng, J., Liu, J., Kuang, H. & Wang, J. A fully automated multimodal mri-based multi-task learning for glioma segmentation and idh genotyping. *IEEE Transactions on Med. Imaging* **41**, 1520–1532 (2022).
27. Graham, S. et al. One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Med. Image Analysis* **83**, 102685 (2023).
28. Zhou, Y. et al. Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. *Med. image analysis* **70**, 101918 (2021).
29. Dong, X. et al. Cross-domain mutual-assistance learning framework for fully automated diagnosis of primary tumor in nasopharyngeal carcinoma. *IEEE Transactions on Med. Imaging* **43**, 3676–3689 (2024).
30. Li, H. B. et al. The brain tumor segmentation (brats) challenge 2023: Brain mr image synthesis for tumor segmentation (brasyn) (2024). [2305.09011](https://arxiv.org/abs/2305.09011).
31. Havaei, M., Guizard, N., Chapados, N. & Bengio, Y. Hemis: Hetero-modal image segmentation. In *International conference on medical image computing and computer-assisted intervention*, 469–477 (Springer, 2016).
32. Varsavsky, T., Eaton-Rosen, Z., Sudre, C. H., Nachev, P. & Cardoso, M. J. Pimms: permutation invariant multi-modal segmentation. In *International Workshop on Deep Learning in Medical Image Analysis*, 201–209 (Springer, 2018).
33. Zhang, Z., Liu, Q. & Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote. Sens. Lett.* **15**, 749–753 (2018).
34. Luo, L. et al. A large model for non-invasive and personalized management of breast cancer from multiparametric mri. *Nat. Commun.* **16**, 3647 (2025).
35. Zhang, Z. et al. Unimrisegnet: Universal 3d network for various organs and cancers segmentation on multi-sequence mri. *IEEE J. Biomed. Heal. Informatics* (2024).
36. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
37. Tang, Y. et al. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20730–20740 (2022).
38. Clark, K. et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *J. digital imaging* **26**, 1045–1057 (2013).

39. Glorot, X., Bordes, A. & Bengio, Y. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323 (JMLR Workshop and Conference Proceedings, 2011).
40. Ulyanov, D., Vedaldi, A. & Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016).
41. Maier-Hein, L., Menze, B. et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. *arXiv.org* (2022).

Acknowledgements

This work was supported by the Macao Polytechnic University Grant (RP/FCA-08/2024) and the Shenzhen Medical Research Fund (D2501013).

Author contributions statement

Y.S. and T.T. conceived and supervised the study. Z.Z., L.H., and D.J. designed the methodology and developed the computational framework. Z.Z., D.J., and J.H. implemented the model architecture and conducted the experiments. Z.Z., L.H., and C.Y. analyzed the experimental results and generated visualizations. S.W., M.K., N.K., and R.M. provided clinical expertise and validated the clinical relevance of the findings. Z.Z. performed data collection, curation, quality control, and wrote the manuscript. T.T., R.M., N.K., and M.K. provided critical revisions. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Supplementary

Internal Datasets

Segmentation Datasets

BraTs2018-GLI dataset. The BraTs-2018 dataset (<https://www.kaggle.com/datasets/sanglequang/brats2018>) comprises 285 subjects with multimodal brain MRI scans for brain tumor segmentation. The dataset includes four MRI sequences: T1-weighted (T1w), T1-weighted contrast-enhanced (T1-ce), T2-weighted (T2w), and T2-FLAIR images. Three tumor regions are annotated: whole tumor, tumor core, and enhancing tumor core. All scans were acquired from multiple clinical institutions following standardized protocols. The BraTs-2018 dataset is used as one of the upstream training datasets for segmentation tasks. We use the standard train/validation/test split provided by the challenge organizers.

HNTS-MRG-24 dataset. The HNTS-MRG-24 dataset (<https://zenodo.org/records/11199559>) comprises 150 subjects with head and neck CT scans for tumor segmentation. The dataset focuses on T2-weighted sequences and provides binary segmentation labels for tumor regions. The images were acquired from patients with various head and neck malignancies using different CT scanners and acquisition protocols. This dataset presents unique challenges due to the complex anatomy of the head and neck region and the heterogeneous appearance of tumors in this area. The HNTS-MRG-24 dataset is used for upstream training in the segmentation framework. We randomly select 75%/5%/20% of the 150 cases for training/validation/testing.

ISPY dataset. The ISPY dataset (<https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=101942541>) comprises 163 subjects with breast DCE-MRI (Dynamic Contrast-Enhanced MRI) scans for tumor segmentation. The dataset provides detailed annotation for breast tumor regions with single-class labeling. The images were acquired from patients undergoing neoadjuvant chemotherapy treatment, allowing for longitudinal assessment of treatment response. The DCE-MRI protocol captures the dynamic enhancement patterns of breast tumors, which is crucial for accurate tumor delineation. The ISPY dataset is used as one of the upstream training datasets. We follow a standard 75%/5%/20% split for training/validation/testing phases.

FedBca dataset. The FedBca dataset (<https://zenodo.org/records/13622759>) comprises 221 subjects with bladder T2-weighted MRI scans for tumor segmentation. The dataset provides single-class annotation for bladder tumor regions. The images represent a diverse collection of bladder cancer cases with varying tumor sizes, locations, and morphologies. All scans were acquired using standardized T2-weighted MRI protocols from multiple clinical centers. The dataset presents challenges related to bladder wall delineation and tumor boundary identification. The FedBca dataset is used for upstream training in our segmentation pipeline. We randomly allocate 75%/5%/20% of the cases for training/validation/testing respectively.

AMOS dataset. The AMOS dataset (<https://zenodo.org/records/7155725#.Y00OC0xBztM>) comprises 60 subjects with abdominal organ segmentation. The dataset provides detailed annotation for 13 different abdominal organ structures including liver, kidneys, spleen, pancreas, stomach, gallbladder, and other major organs. Unlike tumor-focused datasets, AMOS emphasizes multi-organ segmentation in the abdominal cavity. The images were acquired using standardized CT protocols from multiple clinical institutions. The AMOS dataset is used for upstream training with organ segmentation objectives. We employ a 75%/5%/20% data split for training/validation/testing.

PICAI dataset. The PICAI dataset (https://github.com/DIAGNijmegen/picai_labels) comprises 220 subjects with prostate cancer segmentation using multimodal MRI sequences including T2-weighted, ADC (Apparent Diffusion Coefficient), and DWI (Diffusion-Weighted Imaging). The dataset provides single-class annotation for prostate cancer regions. The multimodal approach combining anatomical and functional imaging sequences enables comprehensive characterization of prostate cancer lesions. The images were acquired from multiple institutions using standardized PI-RADS protocols. The PICAI dataset is used as one of the upstream training datasets for cancer segmentation. We follow the standard 75%/5%/20% split for training/validation/testing.

ProstateX dataset. The ProstateX dataset (https://link.csdn.net/?from_id=133825318&target=https%3A%2F%2Fwiki.cancerimagingarchive.net%2Fpages%2Fviewpage.action%3FpageId%3D23691656) comprises 220 subjects with prostate imaging for both organ and tumor segmentation tasks. The dataset utilizes T2-weighted and ADC sequences and provides binary class annotation covering both normal prostate tissue and tumor regions. This dual-target approach allows for comprehensive prostate analysis including both anatomical structure identification and pathological region detection. The images were acquired from patients with suspected or confirmed prostate cancer using standardized multiparametric MRI protocols. The Prostate-X dataset is used for upstream training with dual segmentation objectives. We randomly select 75%/5%/20% of the cases for training/validation/testing.

Liver-HCC dataset. The Liver-HCC dataset (<https://atlas-challenge.u-bourgogne.fr/dataset>) comprises 60 subjects with liver imaging focusing on both organ and tumor segmentation. The dataset uses T1-weighted contrast-enhanced (T1-ce) sequences and provides binary class annotation for liver parenchyma and hepatocellular carcinoma regions. The images were acquired from patients with confirmed HCC diagnosis using standardized contrast-enhanced MRI protocols. The dataset captures various phases of contrast enhancement, which is essential for accurate HCC identification and characterization. The Liver-HCC dataset is used for upstream training in our segmentation framework. We employ a 75%/5%/20% data split for training/validation/testing.

NPC dataset. The NPC (Nasopharyngeal Carcinoma) dataset (<https://zenodo.org/records/10900202>) comprises 277 subjects with head and neck imaging for tumor segmentation. The dataset utilizes multimodal sequences including T1-weighted (T1w), T1-weighted contrast-enhanced (T1-ce), and T2-weighted (T2w) images. Single-class annotation is provided for nasopharyngeal carcinoma regions. The multimodal approach captures different tissue characteristics and contrast enhancement patterns crucial for NPC delineation in the complex nasopharyngeal anatomy. The NPC dataset is used as one of the upstream training datasets. We follow a standard 75%/5%/20% split for training/validation/testing.

Colorectal cancer dataset. The Colorectal cancer dataset (<https://www.imagecomputing.org/isicdm2023/cn/index.html#/Challenge>) comprises 150 subjects with colorectal CT scans for tumor segmentation. The dataset focuses on T2-weighted sequences and provides single-class annotation for colorectal tumor regions. The images represent a diverse collection of colorectal cancer cases with varying tumor locations, sizes, and morphological characteristics throughout the colon and rectum. The dataset presents unique challenges due to the complex anatomy of the colorectal region and the need to distinguish tumor tissue from normal bowel wall and adjacent structures. The Colorectal cancer dataset is used for upstream training. We randomly allocate 75%/5%/20% of the cases for training/validation/testing.

Classification Datasets

BreastDM dataset. The BreastDM dataset (<https://github.com/smallboy-code/Breast-cancer-dataset?tab=readme-ov-file>) comprises 232 subjects (147 malignant and 85 benign) with breast DCE-MRI scans for tumor classification. The dataset utilizes Dynamic Contrast-Enhanced MRI sequences and provides binary classification labels for breast tumor characterization. The images capture the dynamic enhancement patterns of breast lesions, which are crucial for distinguishing between benign and malignant tumors. All scans were acquired using standardized breast MRI protocols from multiple clinical centers. The AMBL dataset is used for downstream classification tasks to evaluate transfer learning capabilities. We randomly select 75%/5%/20% of the cases for training/validation/testing.

BraTs-2018 dataset. The BraTs-2018 dataset (<https://www.kaggle.com/datasets/sanglequang/brats2018>) comprises 285 subjects with multimodal brain MRI scans for tumor classification. The dataset includes four MRI sequences: T1-weighted (T1w), T1-weighted contrast-enhanced (T1-ce), T2-weighted (T2w), and T2-FLAIR images. Binary classification labels are provided for brain tumor type classification. The multimodal approach enables comprehensive characterization of brain tumor features across different imaging contrasts. The BraTs-2018 dataset is used for downstream classification tasks to

assess the generalization ability of pre-trained models. We follow the standard data split provided by the challenge organizers.

LLD-MMRI2023 dataset. The LLD-MMRI2023 dataset (<https://github.com/LMMEng/LLD-MMRI2023>) comprises 314 subjects with liver imaging for tumor classification. The dataset utilizes Portal Venous (PV) and Arterial Portal Venous (APV) phase sequences, providing binary classification labels for liver lesion characterization. The dual-phase imaging approach captures the characteristic enhancement patterns of liver lesions, which is essential for accurate diagnosis. All images were acquired using standardized liver MRI protocols with contrast enhancement. The LLD-MMRI2023 dataset is used for downstream classification tasks. We randomly allocate 75%/5%/20% of the cases for training/validation/testing.

FedBca dataset. The FedBca dataset (<https://zenodo.org/records/13622759>) comprises 275 subjects with bladder T2-weighted MRI scans for tumor classification. The dataset provides binary classification labels for bladder cancer characterization. The images represent a diverse collection of bladder cases including both malignant and benign lesions with varying morphological characteristics. All scans were acquired using standardized T2-weighted MRI protocols optimized for bladder imaging. The FedBca dataset is used for downstream classification tasks to evaluate model performance on bladder pathology. We follow a 75%/5%/20% split for training/validation/testing.

External Datasets

Private-Liver dataset. The Private-Liver dataset comprises 506 subjects with hepatobiliary phase (HBP) MRI scans for liver organ segmentation. This dataset focuses on the hepatobiliary system and provides annotations for liver organ segmentation. All scans were acquired following clinical protocols for hepatobiliary imaging. The dataset is used for evaluating model generalization capabilities on liver segmentation tasks.

BrTS2025-SSA dataset. The BrTS2025-SSA dataset (<https://www.synapse.org/Synapse:syn64153130>) contains 60 multimodal brain MRI scans for brain tumor segmentation. The dataset includes four MRI sequences: T1-weighted (T1w), T1-weighted contrast-enhanced (T1c), T2-weighted (T2w), and T2-FLAIR images. Brain tumor regions are annotated for segmentation analysis. All scans were collected from multiple clinical institutions following standardized neuroimaging protocols. The dataset serves as an external test set for evaluating model performance on brain tumor segmentation.

NACT dataset. The NACT dataset (<https://www.ub.edu/mama-mia/>) consists of 64 breast DCE-MRI scans for breast tumor analysis. The dataset utilizes Dynamic Contrast-Enhanced MRI sequences to capture tumor characteristics in breast tissue. All scans were acquired using standardized breast MRI protocols from clinical centers. This dataset is employed for assessing model performance on breast tumor analysis tasks.

LLD-MMRI2025 dataset. The LLD-MMRI2025 dataset (<https://github.com/LMMEng/LLD-MMRI2023>) includes 104 breast MRI scans for breast tumor analysis, featuring two contrast phases: portal venous phase (PV) and arterial phase (APV). The multi-phase imaging provides complementary information for tumor characterization in breast tissue. All scans were obtained following clinical breast MRI protocols. The dataset is utilized for evaluating model generalization on multi-phase breast MRI analysis.

Unseen Datasets

MSD Hippocampus dataset. The MSD Hippocampus dataset (<https://drive.google.com/drive/folders/1HqEgzS8BV2c7xYNrZdEAnrHk7osJJ--2>) contains 260 T1-weighted (T1w) MRI scans for hippocampus organ segmentation. This dataset specifically focuses on the hippocampus structure and provides detailed annotations for segmentation evaluation. All scans were acquired following standardized neuroimaging protocols. The dataset serves as a completely unseen test set to evaluate model generalization capabilities and transfer learning performance on unseen data domains for hippocampus segmentation tasks.

Table 1. Overview of comprehensive dataset

Dataset	Position	Clinical Target	Sequence	Cls	Subjects	Task
Internal Datasets						
BraTS2018-GLI	Brain	Tumor	T1w, T1c, T2w, FLAIR	3	285	segmentation, classification
HNTS-MRG-24	Head & Neck	Tumor	T2w	2	150	segmentation
ISPY	Breast	Tumor	DCE	1	163	segmentation
FedBca	Bladder	Tumor	T2w	1	221	segmentation, classification
AMOS	Abdomen	Organ	Unknown	13	60	segmentation
PICAI	Prostate	Tumor	T2w, ADC, DWI	1	220	segmentation
Prostate-X	Prostate	Organ & Tumor	T2, ADC	2	220	segmentation
ATLAS	Liver	Organ & Tumor	T1c	2	60	segmentation
NPC	Head & Neck	Tumor	T1w, T1c, T2w	1	277	segmentation, classification
Colorectal cancer	Colorectal	Tumor	T2w	1	150	segmentation
BreaDM	Breast	Tumor	DCE	2	162	classification
LLD-MMRI2023	Liver	Tumor	AP, VP	2	314	classification
External Datasets						
Private-Liver	Liver	Organ	Hepatobiliary Phase (HBP)	1	506	segmentation
BrTS2025-SSA	Brain	Tumor	T1w, T1c, T2w, FLAIR	3	60	segmentation
NACT	Breast	Tumor	DCE	1	64	segmentation
LLD-MMRI2025	Breast	Tumor	AP, VP	2	104	classification
Unseen Datasets						
MSD Hippocampus	Hippocampus	Organ	T1w	2	260	segmentation

Note: For each dataset with n sequences, we evaluated all $2^n - 2$ possible incomplete combinations, excluding only the complete sequence scenario (all sequences present) and the null scenario (no sequences present). HBP: Hepatobiliary Phase; AP: Arterial Phase; VP: Venous Phase; T1w: T1-weighted; T1c: T1-weighted contrast-enhanced; T2w: T2-weighted; FLAIR: Fluid Attenuated Inversion Recovery; DCE: Dynamic Contrast-Enhanced; ADC: Apparent Diffusion Coefficient; DWI: Diffusion Weighted Imaging.

Table 2. Segmentation Performance Comparison between our model with baseline models

Method	Inference	Organs				Tumors				Average			
		DSC (%)	IoU (%)	HD95 (mm)	ASD (mm)	DSC (%)	IoU (%)	HD95 (mm)	ASD (mm)	DSC (%)	IoU (%)	HD95 (mm)	ASD (mm)
ResUnet-specific	Complete-sequence	72.48	61.21	12.65	2.60	56.94	45.16	22.15	9.30	64.71	53.19	17.40	5.95
Early Fuse	Complete-sequence	79.90	69.79	8.82	1.93	59.58	47.68	16.45	7.12	69.74	58.74	12.64	4.53
Late Fuse	Complete-sequence	78.32	68.06	13.47	3.68	57.65	45.94	22.52	10.07	67.99	57.00	18.00	6.88
SAM-Med3D	Complete-sequence	81.15	71.36	6.72	1.52	59.48	47.49	17.79	7.12	70.32	59.43	12.26	4.32
UniMRISegNet	Complete-sequence	81.36	71.64	7.13	1.47	60.78	49.30	15.78	6.81	71.07	60.47	11.46	4.14
MRICombo	Complete-sequence	83.90	75.01	5.08	0.82	63.30	51.57	12.35	4.64	73.60	63.29	8.72	2.73

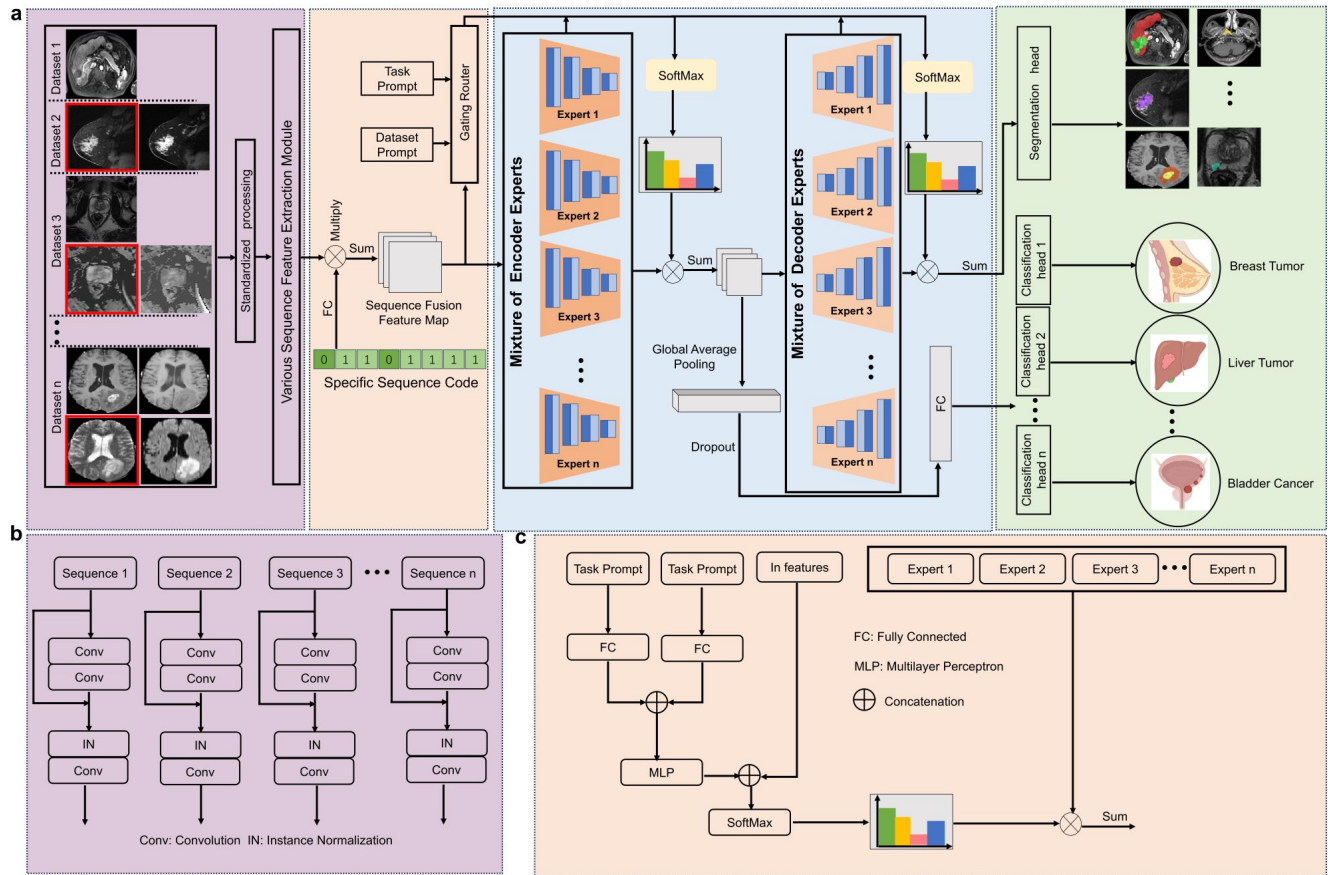


Figure 1. Architectural overview of MRICombo. a, Overall network schematic. b, The various sequence feature extraction module. c, Detailed design of the gating router.