Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

OrganSegBench: A Comprehensive Multi-Organ Benchmark for Segmentation Foundation Models with a Practical Synergy Pathway to Clinical Application

Chengyan Wang

wangcy@fudan.edu.cn

Human Phenome Institution, Fudan University https://orcid.org/0000-0002-8890-4973

Qing Li

Human Phenome Institution, Fudan University

Yizhe Zhang

School of Computer Science and Engineering, Nanjing University of Science and Technology

Xin Guo

Shanghai Academy of Artificial Intelligence for Science https://orcid.org/0009-0003-3414-8564

Haosen Zhang

Bioengineering Department and Imperial-X, Imperial College London

Yan Li

Ruijin Hospital, Shanghai Jiaotong University School of Medicine

Mo Yang

Fudan University https://orcid.org/0009-0004-8565-7235

Yajing Zhang

GE Healthcare

Mengting Sun

Human Phenome Institute and Shanghai Pudong Hospital, Fudan University

Longyu Sun

School of Mechanics and Engineering Science, Peking University

Haoyang Zhang

Human Phenome Institute and Shanghai Pudong Hospital, Fudan University

Junhong Liu

Human Phenome Institute and Shanghai Pudong Hospital, Fudan University

Shuo Wang

Fudan University https://orcid.org/0000-0002-2947-8783

Article

Keywords: Segmentation foundation models, Medical imaging, Benchmark evaluation, Ensemble learning

Posted Date: October 21st, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7844536/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.

OrganSegBench: A Comprehensive Multi-Organ Benchmark for Segmentation Foundation Models with a Practical Synergy Pathway to Clinical Application

Qing Li[1], Yizhe Zhang[2,*], Xin Guo[3], Haosen Zhang[4], Yan Li[5], Mo Yang[1], Yajing Zhang[6], Mengting Sun [1], Longyu Sun[7], Haoyang Zhang[1], Junhong Liu[1], Shuo Wang[8,*], Chengyan Wang[1,*]

[1]Human Phenome Institute and Shanghai Pudong Hospital, Fudan University, Shanghai, China
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, China
[3]Shanghai Academy of Artificial Intelligence for Science, Shanghai, China
[4]Bioengineering Department and Imperial-X, Imperial College London, London, UK
[5]Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[6]GE Healthcare, Beijing, China
[7]School of Mechanics and Engineering Science, College of Engineering, Peking University, Beijing, China
[8]Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, Shanghai, China
* presents the corresponding author

Correspondence to:
Chengyan Wang, Ph.D.
Fudan University
Address: 825 Zhangheng Road, Pudong New District, Shanghai, 201203
Phone: 86-021-3420-8697
E-mail: wangcy@fudan.edu.cn

Abstract

Segmentation Foundation Models (SFMs), despite their success in general computer vision, remain suboptimal for medical imaging, where clinical requirements like fairness and robustness are paramount. Existing benchmarks fail to address these needs and often rely on datasets with potential training data leakage. To bridge this gap, we introduce "OrganSegBench", a comprehensive benchmark evaluation framework built on a new and high quality data resource from 701 subjects with 16 annotated organs and detailed demographics. Using this resource, we systematically evaluated six state-of-the-art SFMs across five key dimensions: accuracy, generalization, robustness, fairness, and clinical utility. Our analysis uncovers a fundamental trade-off: the most accurate models are consistently the least fair, with no single model achieving excellence across all dimensions. To resolve this dilemma, we propose two ensemble strategies: training-free fusion and multi-source knowledge distillation. Notably, both approaches decisively outperformed every individual SFM across all evaluation dimensions, resolving the accuracy-fairness trade-off. These findings expose the inherent limitations of current monolithic SFMs and establish principled model synergy as a practical and superior pathway toward building safe, equitable, and clinically robust AI.

Keywords: Segmentation foundation models, Medical imaging, Benchmark evaluation, Ensemble learning

Main

Medical foundation models hold immense value[1,2], with imaging, a crucial modality in the medical domain, providing unique and comprehensive non-invasive disease assessments. Its paramount challenge lies in achieving fully automated segmentation of multiple organs or lesions, urging the development of foundation models for effective target segmentation. Segmentation foundation models, epitomized by the Segment Anything Model (SAM), have reshaped the landscape of computer vision[3]. Pre-trained on vast, heterogeneous datasets, these models learn powerful feature representations that generalize effectively to new data within familiar tasks, even from unseen datasets[4,5]. Particularly, certain prompt-based foundation models (e.g. SAM) can achieve effective segmentation without any task-specific fine-tuning, demonstrating remarkable zero-shot generalization capabilities[6–9]. This paradigm shift promises to democratize artificial intelligence (AI) by reducing the reliance on costly expert annotations and accelerating the development of adaptable AI systems[10,11]. Their prompt-driven design—using inputs like bounding boxes (bboxes) or text prompts to specify a target—offers a flexible and intuitive interface for complex segmentation tasks, hinting at transformative potential in specialized fields[12,13].

However, the translation of this success into the high-stakes domain of medical imaging remains a formidable challenge[14,15]. The inherent complexity, subtle inter-patient variability, and distinct image properties of clinical data cause generalist models like SAM to yield suboptimal and unstable performance[7]. Domain-specific adaptations, such as MedSAM[16], or advanced 3D architectures like SAT[17] and TotalSegmentator[18], improve accuracy but expose deeper challenges, most notably algorithmic fairness[19–21]. Performance biases across demographic subgroups have been observed, posing a significant barrier to the safe and equitable deployment of AI in clinical practice, where health equity is paramount[22–24].

This translational gap is exacerbated by the limitations of current benchmarking practices. Many evaluations rely on public datasets from multiple sources. Those datasets, while diverse in patient demographics, disease presentations, and imaging modalities, suffer from inconsistent data quality, various annotation standards, and a lack of metadata comprehensiveness[25,26]. These inconsistencies undermine evaluation validity and hinder fairness analyses. Moreover, most public datasets introduce the risk of data leakage[27], as foundation models' pre-training corpora may

inadvertently contain evaluation data, leading to inflated performance estimates. Compounding this issue, most benchmarks remain narrowly focused on accuracy metrics such as the Dice Similarity Coefficient (DSC), neglecting clinically vital dimensions including robustness—reliable performance under challenging conditions—and fairness across diverse patient populations[28]. The result is an incomplete and often overly optimistic assessment of clinical readiness.

To tackle these critical limitations, we introduce a comprehensive framework named "OrganSegBench", aiming to propel foundation model technology from bench to bedside, bridging the gap between research and practical clinical applications (**Fig. 1**). Our work is grounded in a new and high quality dataset from 701 subjects, featuring 16 annotated anatomical structures and corresponding demographic metadata per subject (**Supplementary Fig. 1,2, Supplementary Table 1**). We utilized the data resource from PhenoBank (https://www.phenobank.org.cn/), which was publicly released for the first time by our institution. This dataset ensures annotation consistency and eliminates the risk of data leakage. Using this OrganSegBench framework, we conducted a multi-dimensional evaluation of six prominent SFMs, assessing their accuracy, generalization, robustness, demographic fairness, and clinical utility. Furthermore, we introduce two ensemble strategies, training-free fusion and multi-source knowledge distillation, to improve SFM performance and cost-effective clinical deployment. Our results reveal not only the fundamental constraints of current models but also that ensemble-based approaches consistently deliver superior performance across all five evaluation dimensions. This work establishes both a new benchmark and deployment paradigm for advancing the next generation of medical segmentation models.

Results

Overview

As illustrated in **Fig.1A-F**, OrganSegBench provides a framework for evaluating diverse segmentation foundation models (SFMs). Its adaptable design accommodates heterogeneous SFM frameworks, such as bbox-prompted, text-prompted, and prompt-free approaches—enabling standardized benchmarking across imaging modalities and laying the groundwork for future online platform development. The framework assesses SFM performance across five critical dimensions:

accuracy, generalization, robustness, fairness, and clinical utility. Comparative analysis revealed that bbox-prompted SFMs (SAM[3], MedSAM[16], SAM2[29], MedSAM2[30]) consistently outperformed text-prompted (SAT[31]) and prompt-free (TotalSeg[18]) models (**Supplementary Table 2**) in accuracy, generalization, robustness, and clinical utility. Conversely, SAT and TotalSeg achieved superior fairness across attributes of age, gender and Body Mass Index (BMI), an important consideration for equitable healthcare deployment (**Fig.1H**). To probe spatial heterogeneity in segmentation errors, we further developed a novel visualization method that maps anatomical region–specific error distributions into organ point cloud. This method offers radiologists targeted quality-control insights and guides researchers in developing task-specific optimizations for critical anatomical areas. Finally, we implemented two ensemble strategies. The first is a training-free fusion strategy that integrates predictions from all six SFMs with simple logical operations. The second strategy proposes *multi-source* knowledge distillation to train a single, specialized model from their combined knowledge. Both approaches consistently surpassed the performance of any individual model, demonstrating a lightweight yet effective pathway toward clinically robust, equitable, and high-performing medical AI.

Individual foundation models show performance heterogeneity

As shown in **Fig.2** and **Supplementary Table 3-6**, our investigation reveals that the clinical potential of current SFMs is undermined by profound performance heterogeneity. We found a distinct performance hierarchy among the six models (**Fig. 2A, C-E**), where bbox-prompted architectures (DSC of 0.77, 0.79, 0.66, and 0.76 for SAM, SAM2, MedSAM, MedSAM2, respectively) consistently achieved higher overall accuracy than text-prompted (DSC of 0.61 for SAT) or prompt-free (DSC of 0.58 for TotalSeg) approaches. While these models excelled at segmenting large, well-demarcated organs (e.g. lung, liver, kidney, spleen), their performance degraded significantly on smaller or more complex targets (e.g. pancreas, vessels). This weakness was particularly acute for SAT and TotalSeg, which went through a major breakdown (DSC < 0.4) in segmenting cardiac substructures. The similar results were observed using 95% Hausdorff Distance(HD95) for evaluation (**Fig. 2B, F-H**). Notably, the MedSAM outperformed other SFMs in vessel segmentation (**Fig. 2H**, ranked by mean HD95 among vessels segmentation: MedSAM[4.7] < SAM[6.0] < MedSAM2[6.1] < SAM2[7.3] < TotalSeg [29.4] < SAT[31.7]).

Counter-intuitively, domain specialization did not guarantee superior performance. The medically-focused MedSAM was consistently outperformed by the generalist SAM (mean DSC[0.66], HD95[9.6] of MedSAM vs. mean DSC[0.77], HD95[8.6]) in large organ segmentation, challenging the prevailing assumption that domain-specific fine-tuning is universally beneficial. Additionally, while the segmentation performance of other SFMs exhibited a positive correlation with the degree of consistency with radiologists' annotations, MedSAM demonstrated a negative correlation in large-organ segmentation and showed suboptimal performance on samples considered to be of low segmentation difficulty (**Fig. 2I, J**). This suggests it can erode the robust generalizability learned from vast, diverse datasets. To further dissect performance, we visualized the spatial distribution of errors for five representative organs (**Fig. 3A-E**). This spatial analysis revealed distinct, model-specific limitation modes: bbox-prompted models tended to produce errors in the core of organs, whereas bbox-free models like SAT and TotalSeg were more prone to errors at anatomical boundaries. Notably, all models struggle with challenging sub-regions, such as the renal medulla, highlighting areas that may require expert oversight (**Fig. 3C**). These findings underscore that a single accuracy score can mask diverse and critical underlying error patterns, as even the top-performing models, SAM and SAM2, exhibited substantial performance variability across different anatomical sub-regions.

Performance on challenging cases reveals clinical brittleness

A model's true clinical viability is determined not by its average performance, but by its robustness in difficult, real-world cases[32]. To simulate this, we stress-tested each model on the most challenging 5% of cases for each organ—a sub-cohort where diagnostic accuracy is most critical. As shown in **Fig. 4**, in these high-stake scenarios, the superior robustness of bbox-prompted models became evident; they largely maintained their segmentation quality with only minor degradation. In stark contrast, SAT and TotalSeg collapsed catastrophically, particularly on cardiac segmentation (**Fig. 4A, C-E**). The similar results were observed using HD95 for performance evaluation on challenging cases(**Fig. 4B, F-H**). Notably, the performance of SAM2 was decreased significantly in vessel segmentation(**Fig. 4H**). This wbrittleness exposes a critical flaw: models that appear adequate on average may be dangerously unreliable when faced with atypical anatomy or lower-quality scans. This finding underscores that conventional evaluation

metrics are insufficient for assessing clinical safety[28] and that robustness to challenging cases must be a primary benchmark for any model intended for medical use.

A fundamental trade-off exists between model accuracy and demographic fairness

Beyond accuracy and robustness, the ethical deployment of medical AI hinges on fairness. Our analysis of model performance across demographic subgroups—gender, age, and BMI[33,34]—uncovered a deeply concerning trade-off: the most accurate models were consistently the least fair.

Shown in **Fig. 5**, SAM and SAM2, despite leading in accuracy (DSC), displayed significant, systematic biases, particularly regarding gender(**Fig. 5A, G**), where they performed better on male subjects for the majority of organs (**Fig. 5H**). Conversely, SAT, one of the weaker models on accuracy metrics, demonstrated markedly greater fairness. Similarly, bbox-prompted models exhibited substantial fairness issues across age and BMI attributes (**Fig. 5B, C**). For instance, SAM and SAM2 achieved higher accuracy in older age groups and subgroups with higher BMI (**Fig. 5H**). Notably, we found that under the HD95 metric, these models showed improved fairness relative to the DSC metric with respect to gender and age, indicating consistently equitable performance in segmenting organ boundaries. This inverse relationship between accuracy and fairness exposes a critical dilemma for clinical translation: the architectural features or data priors that drive high accuracy may simultaneously amplify systemic biases. These biases were complex and attribute-dependent, with no single SFM achieving equitable performance across all demographic groups. Additionally, the divergence between HD95 and DSC reveals that fairness is metric-dependent: models can exhibit improved equity in boundary delineation while showing biases in volumetric overlap. This underscores the necessity of context-specific fairness evaluations aligned with clinical priorities. This work quantifies the tangible risk that deploying current state-of-the-art models could inadvertently perpetuate or even exacerbate existing health inequities.

Contradictory performance profiles limit the clinical utility of any single model

To connect abstract metrics with practical use, we assessed the clinical utility of each SFM on two common downstream tasks: cardiac function analysis and multi-organ volumetry[35,36]. The results

crystallized the core problem: no single model offers a universally reliable solution due to contradictory performance profiles.

For quantifying cardiac phenotypes, such as ejection fraction, the bbox-prompted models' high accuracy in cardiac segmentation translated directly to superior utility (**Table 1**). However, this performance hierarchy inverted when assessing multi-organ volumetric accuracy for non-cardiac structures. In this task, SAT and TotalSeg—despite their cardiac failures—outperformed the more accurate SAM-family models. Their competent segmentation of large abdominal organs was sufficient to make them the better choice for general volumetry. This paradox highlights an untenable situation for clinical practice: one must choose between a model that excels at cardiac analysis and another that is better for abdominal volumetry, making a unified, efficient workflow impossible with any single SFM.

A synergistic ensemble of models overcomes individual limitations

Given the complementary strengths and weaknesses of individual SFMs, we hypothesized that a synergistic ensemble could transcend the limitations of any individual model. We validated this by evaluating five training-free ensemble strategies and three multi-source knowledge distillation methods.

As shown in **Fig. 6**, the results were clear: both ensemble approaches consistently outperformed every individual model. Among training-free fusion strategies, a simple average (AVG) fusion of all six models dominated every individual SFM. The AVG fusion achieved superior accuracy (mean DSC 0.83 vs. 0.79 for the best individual model, **Fig. 6A**), demonstrated greater robustness on challenging cases, and offered the most consistent generalization across organs (**Fig. 6C, E**). Crucially, it resolved the accuracy-fairness trade-off by maintaining high fairness without compromising its top-tier accuracy (**Fig. 6G-I**). Similar results were observed by using HD95 metrics. In terms of clinical utility, it delivered the most accurate cardiac measurements while remaining highly competitive for multi-organ volumetry (**Table 1**).

Multi-source knowledge distillation produced even better results. Distilled student models consistently achieved a new state-of-the-art in accuracy, with the distilled PUNet model reaching a mean DSC of 0.86. This specialist approach also delivered superior generalization, evidenced by

a lower standard deviation in accuracy across organs (0.07 vs. 0.11 for the best individual model), while simultaneously preserving and enhancing fairness. More detailed results of ensemble learning strategies in organ, vessel segmentation performance enhancement can be found in **Supplementary Fig. 3,4**. In summary, both ensemble strategies, despite of their heterogeneous design, produced a clinically more reliable and equitable performance profile than any individual SFM, highlighting a direct path toward more robust real-world utility (**Fig. 6M, N**). This result suggests a paradigm shift: instead of seeking a single all-purpose foundation model, safer, more effective, and more equitable clinical AI may be achieved through principled fusion of diverse models.

Discussion

Our comprehensive benchmarking of SFMs reveals a critical disconnect between their theoretical promise and their practical readiness for clinical deployment[37]. By establishing OrganSegBench, a framework grounded in a meticulously curated, single-institution dataset, we moved beyond simplistic accuracy metrics to conduct a multi-dimensional comprehensive evaluation. The results are sobering: no single state-of-the-art model performs well across all essential clinical relevant dimensions. Most notably, we identified a fundamental and pervasive trade-off between segmentation accuracy and demographic fairness—a dilemma that poses a substantial ethical barrier to clinical adoption. That high accuracy can inadvertently amplify bias underscores the urgent need to reconsider optimization priorities in medical AI.

Furthermore, our analysis challenges several prevailing assumptions. The superior performance of the generalist SAM over the domain-specialized MedSAM suggests that current fine-tuning strategies may be overly restrictive, potentially sacrificing the robust representations learned from diverse, large-scale data[22,38]. Moreover, by visualizing spatial error distributions, we provide a new diagnostic lens for understanding how models fail, revealing distinct error topographies (e.g., central vs. boundary failures) that are invisible to standard metrics like DSC[28]. This shift in perspective moves the field from simply measuring outcomes toward understanding model behavior. By simulating real-world clinical tasks such as quantification of cardiac features and multi-organ volumes, we demonstrate that model performance is highly task-dependent, underscoring the necessity of context-aware model selection and the limitations of relying on

isolated metrics. Our pragmatic, application-oriented evaluation framework, OrganSegBench not only provides a multi-dimensional assessment incorporating critical clinical metrics but is also designed to adaptively integrate future clinical tasks such as disease prediction, surgical planning, and morphological analysis, thereby systematically bridging the gap between algorithmic performance and real-world usability and accelerating the clinical deployment of foundational models.

The most significant conceptual advance of our work is the demonstration that ensemble learning strategies, even a simple, training-free ensemble of existing models can overcome the limitations of any single constituent. The consistent superiority of the AVG fusion and multi-source knowledge distillation across almost all evaluation dimensions offers a powerful paradigm shift. It challenges the dominant trajectory in AI development, which often prioritizes the creation of ever-larger, monolithic models as a generalist for all tasks[39–41]. Notably, the choice between these two ensemble strategies can be tailored to specific application scenarios. The voxel-averaging ensemble requires no prior training and is well-suited for smaller-scale or heterogeneous datasets, enabling rapid and straightforward deployment. In contrast, multi-source distillation leverages learned target-domain knowledge, making it ideal for larger, more consistent datasets where post-training efficiency and task-specific optimization are critical. Together, these approaches underscore the high potential of ensemble strategies for enhancing the overall performance of foundation models. Our findings suggest a more pragmatic and potentially more effective path forward: model synergy over model scale. The future of clinical AI may not rest on a single "super-model," but on the intelligent and principled fusion of diverse, complementary architectures, including Mixture of Experts (MoE), which has proven highly effective in enhancing the performance of foundation models[42].

This work also establishes OrganSegBench not as a static leaderboard, but as a living framework to drive community progress. We acknowledge, however, that our dataset originates from a single institution. This raises important questions about how our findings generalize across different imaging facilities, protocols, and patient populations. Future work must validate these results on multi-institutional data with more cases. Meanwhile, we need to implement stringent quality control measures for multi-center data and maintain uniform standards during segmentation annotation. Additionally, our study focused on zero-shot performance; the dynamics of these trade-

offs might change with model fine-tuning. We plan to address these limitations by developing OrganSegBench into an open online platform, enabling the community to benchmark new models and helping to build a more robust, multi-institutional evidence base. Additionally, since SAM-based models require bounding boxes as input, their performance is not annotation-free but implicitly relies on weak supervision. This introduces a potential dependency on both the quality and consistency of human annotators. To mitigate this limitation in our study, all bounding boxes were generated by multiple radiologists using a semi-automatic tool following established protocols (e.g., MedSAM). This approach ensures high annotation consistency at a lower cost compared to full mask annotation. Nevertheless, future efforts should focus on reducing this human dependency, for example, by incorporating automated bounding box generation derived from prompt-free models, segmentation-prior networks, or generative prompting strategies[43]. Such advancements would minimize anthropic bias and enhance the reproducibility of model performance.

In conclusion, this study provides a necessary reality check on the current state of segmentation foundation models in medicine. The uncritical translation of generalist models to clinical practice carries substantial risk. Even state-of-the-art SFMs exhibit critical weaknesses in robustness and fairness that are obscured by accuracy metrics alone. We reveal a fundamental trade-off between performance and demographic equity, challenging the notion that a single universally superior model is attainable—or even desirable. Crucially, our results point to a constructive way forward: ensemble learning strategies systematically outperform individual models across every dimension, from accuracy to fairness. This principle of synergy over scale marks a paradigm shift for clinical AI, moving away from monolithic models toward intelligent synthesis of complementary systems. By diagnosing current limitations and demonstrating the power of collective approaches, this work provides a roadmap for building the next generation of medical AI—effective, safe, equitable, and ultimately worthy of clinical trust.

Methods

The OrganSegBench Dataset

Data Acquisition and Cohort. To evaluate the performance of SFMs and eliminate potential data leakage from pre-training corpora, we utilized the data resource from PhenoBank (https://www.phenobank.org.cn/) that was publicly released for the first time. This study adhered to the ethical standards of the 1975 Declaration of Helsinki(No.BE1828). The dataset contains multi-omics data from 1,008 individuals. We conducted screening on the complete data comprising images of all organs and ultimately included multi-organ images from 701 individuals in this study. The dataset includes multi-modal imaging scans (Chest-CT, Abdomen-MRI, and Heart-MRI) from a single institution. This dataset ensures consistency in imaging protocols and annotation standards. All participants provided informed consent, and the study was approved by the institutional review board. Key acquisition parameters were standardized as below.

- **Chest-CT:** Scans were acquired with a tube voltage of 100 kV, tube current of 134 mAs, slice thickness of 1.0 mm, FOV of 512 × 512 mm^2, and spatial resolution of 0.68 × 0.68 mm^2. A breath-holding protocol was used for stability. Images were reconstructed using fast cone-beam filtered back projection. Windowing was set to a width/level of 1, 000/ − 650 HU for lung and 350/40 HU for mediastinum.
- **Abdomen-MRI:** Scans of the liver, spleen, and kidneys utilized a two-echo mDIXON-TSE sequence on the axial plane with breath-holding. Key parameters included a slice thickness of 2.5 mm, FOV of 400 × 400 mm^2, spatial resolution of 0.88 × 0.88 mm2, TR/TE of 1.31/2.20 ms, flip angle of 10°, and a SENSE factor of 2.0.
- **Heart-MRI:** Cine imaging with 24 cardiac phases was performed using a segmented TrueFISP sequence. Short-axis (SA) views were acquired with a slice thickness of 8.0 mm, FOV of 400 × 400 mm2, spatial resolution of 0.7 × 0.7 mm^2, and TR/TE of 3.42/1.50 ms.

Further details on imaging protocols and inclusion/exclusion criteria are provided in the Supplementary Materials[44,45].

Manual Annotation. For each of the 701 subjects, a total of 16 anatomical structures (9 large organs and 7 major vessels) were manually segmented to create ground-truth masks. All annotations were performed by five experienced radiologists with over 10 years of experience. The overall segmentation consistence among the annotations from different radiologists are shown in **Supplementary Fig. 1**. The dataset also includes demographic information for fairness

evaluation: gender (289 males, 412 females), age (ranging from 20 to 60 years), and BMI (ranging from 14.8 to 48.4 kg/m2).

Organ Segmentation using SFMs

We benchmarked six prominent SFMs of 3 paradigm (Bbox-prompted, Text-Prompted, Prompt-Free), selected to represent the current diversity in model architecture and prompting strategy. For a given dataset D, let the i-th subject be S_i and their 3D image of a specific modality be $X_{i,\text{mod}}$. The evaluated models were SAM ($\mathcal{M}_{SAM}$), MedSAM ($\mathcal{M}_{MedSAM}$), SAM2 ($\mathcal{M}_{SAM2}$), MedSAM2 ($\mathcal{M}_{MedSAM2}$), SAT ($\mathcal{M}_{SAT}$), and TotalSegmentator ($\mathcal{M}_{TotalSeg}$). All models were used with their original pre-trained weights.

- **Bbox-Prompted Models:** $\mathcal{M}_{SAM}$, $\mathcal{M}_{MedSAM}$, $\mathcal{M}_{SAM2}$, and $\mathcal{M}_{MedSAM2}$ are 2D models requiring prompts. 3D images X_i were processed as a sequence of 2D slices $x_{i,j}$. Based on prior work showing the superiority of bounding box prompts, we used the ground-truth bbox $B_{i,j,o}$ as the prompt for segmenting organ o in slice j. The resulting segmentation mask for the j-th slice is denoted as $\widehat{Y_{i,j,o}^{N}} = \mathcal{M}_{\mathcal{N}}(x_{i,j}, B_{i,j,o})$, where $N \in \{1, 2, 3, 4\}$ corresponds to the four models.
- **Text-Prompted Model**: $\mathcal{M}_{SAT}$ is a 3D model that accepts text prompts. To ensure consistency, we used simple prompts consisting of the modality and organ name (e.g., "CT" and "Lung"). The segmentation output for organ o in image $X_{i,\text{mod}}$ is denoted as $\widehat{Y_{i,\text{mod},o}^{\text{SAT}}} = \mathcal{M}_{SAT}(X_{i,\text{mod}}, mod, o)$.We used the MONAI library to restore output dimensions to match the original input.
- **Prompt-Free Model**: $\mathcal{M}_{TotalSeg}$ is a 3D model that requires no external prompts. It was applied directly using its default configuration, with the output defined as $\widehat{Y_{i,\text{mod},o}^{\text{TotalSeg}}} = \mathcal{M}_{TotalSeg}(X_{i,\text{mod}})$.

Multi-Dimensional Performance Evaluation

Accuracy: We evaluated volumetric and boundary accuracy using the Dice Similarity Coefficient (DSC) and the 95$^{\text{th}}$ percentile Hausdorff Distance (HD95), respectively. The average performance for a given model m on an organ o across all samples was recorded as $DSC_{o,m}^{\text{all}}$ and $HD95_{o,m}^{\text{all}}$. We

also analyzed performance correlation with annotation difficulty. For a subset of 30 samples, five radiologists performed repeated annotations. The inter-rater DSC was used to stratify the samples into low, medium, and high annotation difficulty groups, against which model performance was compared. Furthermore, we analyzed the spatial distribution of segmentation errors to identify anatomical regions of failure (details in Supplementary Materials).

Robustness: We assessed model performance on the most challenging cases. For each model m and organ o, we identified the subset of samples corresponding to the lowest 5% DSC values, $S_{o,m}^{\mathrm{DSC},\,5\%}$, and the highest 5% HD95 values, $S_{o,m}^{\mathrm{HD95},\,5\%}$. The average performance on these subsets, $\overline{DSC}_{o,m}^{5\%}$ and $\overline{HD95}_{o,m}^{5\%}$, was calculated to quantify robustness. Performance degradation was measured as the difference between the overall average and this worst-case average.

Fairness: To investigate algorithmic bias, we stratified the cohort by **gender** (male, female), **age** (20-30, 30-40, 40-50, 50-60 years), and **BMI** (< 21, 21-24, >24). For each sensitive attribute, a **one-way analysis of variance** (**ANOVA**) was applied to the DSC and HD95 distributions to test for statistically significant performance differences among subgroups. A non-significant result ($p \geq 0.05$) was interpreted as fairer performance. We also quantified systematic bias by calculating “performance advantage counts" for each subgroup, tallying the number of organs for which a model performed significantly better for that subgroup.

Clinical Utility: We evaluated the applicability of the models for downstream clinical tasks. For cardiac analysis, we computed eight phenotypes (LVEDV, RVEDV, LVESV, RVESV, LVSV, RVSV, LVEF, RVEF) from the model generated masks and calculated the **absolute error (and error rate)** against phenotypes derived from ground-truth masks. For other organs, we calculated the volume from the ground-truth mask (V_{GT}) and the model’s segmentation ($\hat{V}$) to derive the **Volume Error Rate**,$VER = \frac{|V_{\mathrm{GT}} - \hat{V}|}{V_{\mathrm{GT}}}$.

Segmentation Error Distribution Computation

In contrast to the traditional reliance on metrics such as DSC and HD, we analyzed model performance by examining the overall spatial distribution of segmentation errors, which is crucial for downstream tasks and the clinical applicability of SFMs. For the segmentation results derived

from 2D models, we reconstructed them into 3D segmentation outcomes through stacking. To quantify the spatial errors of these 3D segmentation results, we employed the Euclidean distance to the ground truth. Specifically, both the segmentation results and the ground truth were saved in point cloud data format. The set of Euclidean distances from each point in the ground truth point cloud to the closest point in the segmented result point cloud was defined as the spatial distribution of segmentation errors. The process of calculating the averaged point cloud error distribution across the dataset can be found in the **Supplementary Materials**.

Ensemble Learning Strategies

Training free fusion. To mitigate the weaknesses of individual models, we explored five post-processing ensemble strategies to fuse the segmentation masks from all six SFMs. For a given voxel p in the 2D slice mask for organ o from model N, let its value be $\widehat{Y^{N}_{i,J,o}}(p) \in \{0,1\}$. The fused mask voxel is $\widehat{Y^{\text{ensemble}}_{i,J,o}}(p)$.

- **OR/AND**: Logical operations on the six masks. The output voxel is $\widehat{Y^{\text{OR}}_{i,J,o}}(p) = \cup_N \left(\widehat{Y^{N}_{i,J,o}}(p)\right)$ for OR, and $\widehat{Y^{\text{AND}}_{i,J,o}}(p) = \cap_N \left(\widehat{Y^{N}_{i,J,o}}(p)\right)$ for AND.
- **MIN/MAX**: Voxel-count-based selection. Let $C^{N}_{i,j,o} = \sum_p \widehat{Y^{N}_{i,J,o}}(p)$ be the number of segmented pixels for model N. The final mask is chosen from the model with the minimum or maximum count. For MIN, the resulting mask is $\widehat{Y^{\text{MIN}}_{i,J,o}} = \widehat{Y^{N^*}_{i,J,o}}\ where N^* = \arg\min_N C^{N}_{i,j,o}$.
- **Averaging (AVG):** A voxel-level averaging of the six masks followed by thresholding (in Eq. 1).

$$\widehat{Y^{\text{AVG}}_{i,J,o}} = \begin{cases} 1, & \frac{1}{6}\sum_{N=1}^{6} \widehat{Y^{N}_{i,J,o}}(p) \geq 0.5 \\ 0, & \frac{1}{6}\sum_{N=1}^{6} \widehat{Y^{N}_{i,J,o}}(p) < 0.5 \end{cases} \quad (1)$$

The resulting 2D ensemble masks were stacked to reconstruct 3D volumes, which were then evaluated across all five performance dimensions.

Multi-Source Knowledge Distillation. While SFMs can perform multi-target segmentation using a single model as a generalist, we observed significant performance variations across different segmentation targets. To address this limitation, we propose a constructive framework based on Specialist models and ensemble learning mechanisms, termed Multi-Source Knowledge Distillation. This framework comprises two distinct approaches: Deterministic Distillation and Probabilistic Distillation.

Deterministic Distillation. In this approach, the multiple expert masks are aggregated into a single "consensus" mask. This consensus mask serves as the pseudo-label for training a small (relative to the foundation) student model. We evaluate two representative student model architectures: UNet[46] and HSNet[47], where UNet is CNN-based model and HSNet is a Transformer-driven model (using PvT[48]). For each image X_i, the average of masks generated by 6 foundation model is computed as Eq.1. The student model, denoted as S, is trained to predict $\widehat{Y_{i,J,o}^{\mathrm{AVG}}}$ given the input 2D slice image X_i. The loss function is a combination of Intersection over Union (IoU) loss[49] and Binary Cross-Entropy (BCE) loss[50] (in Eq. 2).

$$\mathcal{L}\left(X_i, \widehat{Y_{i,J,o}^{\mathrm{AVG}}}\right) = \sum_{m=1}^{M} \left[\mathcal{L}_{IOU}\left(S_m(X_i), \widehat{Y_{i,J,o}^{\mathrm{AVG}}}\right) + \mathcal{L}_{BCE}\left(S_m(X_i), \widehat{Y_{i,J,o}^{\mathrm{AVG}}}\right)\right] \quad (2)$$

where $S_m(X_i)$ represents the m-th output of the student model and M represents the number of segmentation generalized. For UNet, $M = 1$. For HSNet, $M = 4$, consistent with the original HSNet publication.

For UNet inference, the final prediction is obtained by applying a sigmoid activation to the model's output: $S_{inference}(X_i) = \sigma\big(S(X_i)\big)$. For HSNet, the final prediction is obtained by averaging the M outputs after a sigmoid activation: $S_{inference}(X_i) = \sigma\left(\frac{1}{M}\sum_{m=1}^{M} S_m(X_i)\right)$, where σ is the sigmoid function.

Probabilistic Distillation. This approach employs the Probabilistic U-Net (PUNet), a model designed to accommodate multiple plausible segmentation masks by learning a distribution over possible outputs.

We incorporate a latent space R^N ($N = 6$in our implementation, corresponding to the number of SFMs) to represent segmentation variability. The full model consists of a prior network $P(z|X_i)$,

a posterior network $Q(z|X_i, Y_i)$, and a U-Net backbone. *(a) Prior Network*: Generates a Gaussian distribution over the latent space, conditioned on the input image X_i: $z \sim P(\cdot\,|X_i) = \mathcal{N}\left(\mu_{prior}(X_i; \omega), \Sigma_{prior}(X_i; \omega)\right)$, where ω represents the parameters of the prior network. *(b) Posterior Network*: Generates a Gaussian distribution conditioned on both the input image X_i and a target mask $z \sim Y_i$: $Q(\cdot\,|X_i, Y_i) = \mathcal{N}\left(\mu_{post}(X_i, Y_i; \nu), \Sigma_{post}(X_i, Y_i; \nu)\right)$, where ν denotes the parameters of the posterior network. *(c) Segmentation Generation*: A latent vector z, sampled from the latent space distribution, is combined with features extracted by the U-Net backbone, $f_{U-Net}(X_i; \theta)$, to produce a segmentation mask: $S(X_i, z) = f_{comb}(f_{U-Net}(X_i; \theta), z; \psi)$where θ are the U-Net parameters, f_{comb} represents a combining function (implemented as three sequential 1x1 convolutional layers), and ψ are the parameters of f_{comb}.

During training, for each image X_i, one of the six SFM-generated masks is randomly selected as the target mask. Let K be a discrete uniform random variable taking values in the set$\{1, 2, 3, 4, 5, 6\}$. The target mask Y_i is defined as: $Y_i = \sum_{k=1}^{6} I(K = k) Y_{i,F_k}$, where $I(\cdot)$ denotes the indicator function.

The loss function encourages accurate segmentation and consistency between the prior and posterior distributions (in Eq. 3).

$$\mathcal{L}(X_i, Y_i) = \boldsymbol{E}_{\boldsymbol{z}\sim\boldsymbol{Q}(\cdot|X_i, Y_i)}\left[-\log P_c\left(Y_i \middle| S(X_i, \boldsymbol{z})\right)\right] + \boldsymbol{\beta} \cdot D_{KL}(Q(\boldsymbol{z}|X_i, Y_i) \middle| P(\boldsymbol{z}|X_i)) \qquad (3)$$

where P_c is the likelihood of the correct segmentation, D_{KL} denotes the Kullback-Leibler divergence, and β is a weighting hyperparameter.

At inference time, given an input image X_i, we aim to generate a single, high-quality segmentation mask. This is achieved by leveraging the learned prior distribution $P(z|X_i)$ to sample N (N =10 in our study) latent vectors, $\{z_1, z_2, \ldots, z_N\}$, where $z_n \sim P(\cdot\,|X_i)$ for $n = 1, \ldots, N$. Each latent vector z_n is then used to generate a corresponding segmentation mask $S(X_i, z_n)$. The final segmentation mask, $S_{final}(X_i)$, is obtained by averaging these individual masks:$S_{final}(X_i) = \frac{1}{N}\sum_{n=1}^{N} S(X_i, z_n)$.

Data availability

The datasets used in this study are available from PhenoBank (https://www.phenobank.org.cn/). By submitting a request in the system, full access to the data can be obtained.

Code availability

The code for computation and visualization of segmentation error distributions and training/inferencing of student models for multi-source distillations have be released at https://github.com/lq2233/OrganSegBench.

Funding

This work was supported in part by the Shanghai Municipal Science and Technology Major Project (No.2023SHZDZX02A05), the Shanghai Rising-Star Program (No.24QA2703300), the National Natural Science Foundation of China (No.62331021, 62501166, 62201263), the Natural Science Foundation of Jiangsu Province (No.BK20220949), the International Science and Technology Cooperation Program under the 2023 Shanghai Action Plan for Science (No.23410710400), the National Key R&D Program of China (No.2024YFC3405800, 2024YFF1207500), the Shanghai Sailing Program (No.20YF1402400, 22YF1409300) the Scientific Research Fund Project of Pudong Hospital Affiliated to Fudan University (No.YJJC202409), Shanghai Municipal Education Commission Project for Promoting Research Paradigm Reform and Empowering Disciplinary Advancement through Artificial Intelligence (No.24RGZNC01)

Acknowledgement

We gratefully acknowledge the support from the CFFF platform of Fudan University for providing computational resources for this research.

Clinical trial number

Not applicable.

Ethics statement

Ethical approval was granted by Fudan University (No.BE1828). Written informed consent was obtained from all participants.

Author contributions

Study design: Y.Z.Z., S.W., C.Y.W, Q.L. Data collection: C.Y.W, Q.L. Computation resource supplement: C.Y.W., X.G. Methodology: Y.Z.Z., Q.L. Image processing: Q.L. Model training and optimization: Q.L. Statistical analysis: Q.L. Manuscript preparation: Y.Z.Z., S.W., C.Y.W., Q.L,

H.S.Z., Y.L., M.Y., Y.J.Z., M.T.S., L.Y.S., H.Y.Z., J.H.L. Manuscript review and revision: all authors.

Competing interests

The authors declare no competing interests.

Figures

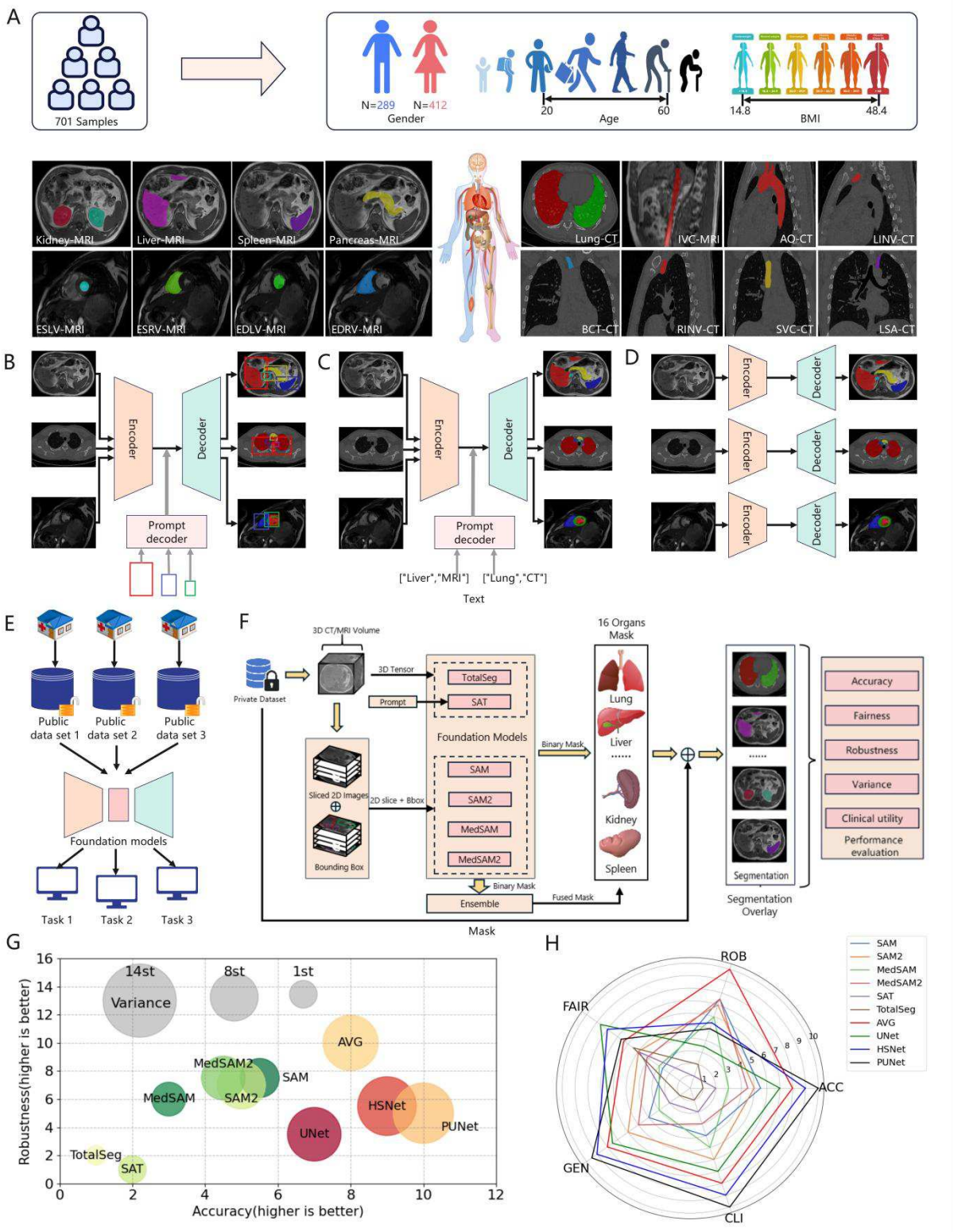


Figure 1 | Overview of the OrganSegBench framework and key findings. A, The private dataset of 701 subjects, annotated for 16 organs/vessels with detailed demographic profiles. **B-D**,

The diverse SFM architectural paradigms evaluated in this study: **(B)** bbox-prompted, **(C)** text-prompted, and **(D)** prompt-free models. **E-F**, The benchmark workflow, where SFMs pre-trained on large-scale data are subjected to **(F)** our five-dimensional evaluation pipeline. **G-H**, A summary of the benchmark results, illustrating **(G)** the multi-dimensional performance trade-offs between individual models and **(H)** the consistent superiority of the AVG fusion strategy and multi-source distillation across evaluation dimensions.

*ACC: Accuracy, ROB: Robustness, FAIR: Fairness, CLI: Clinical Utility, GEN: generalization .

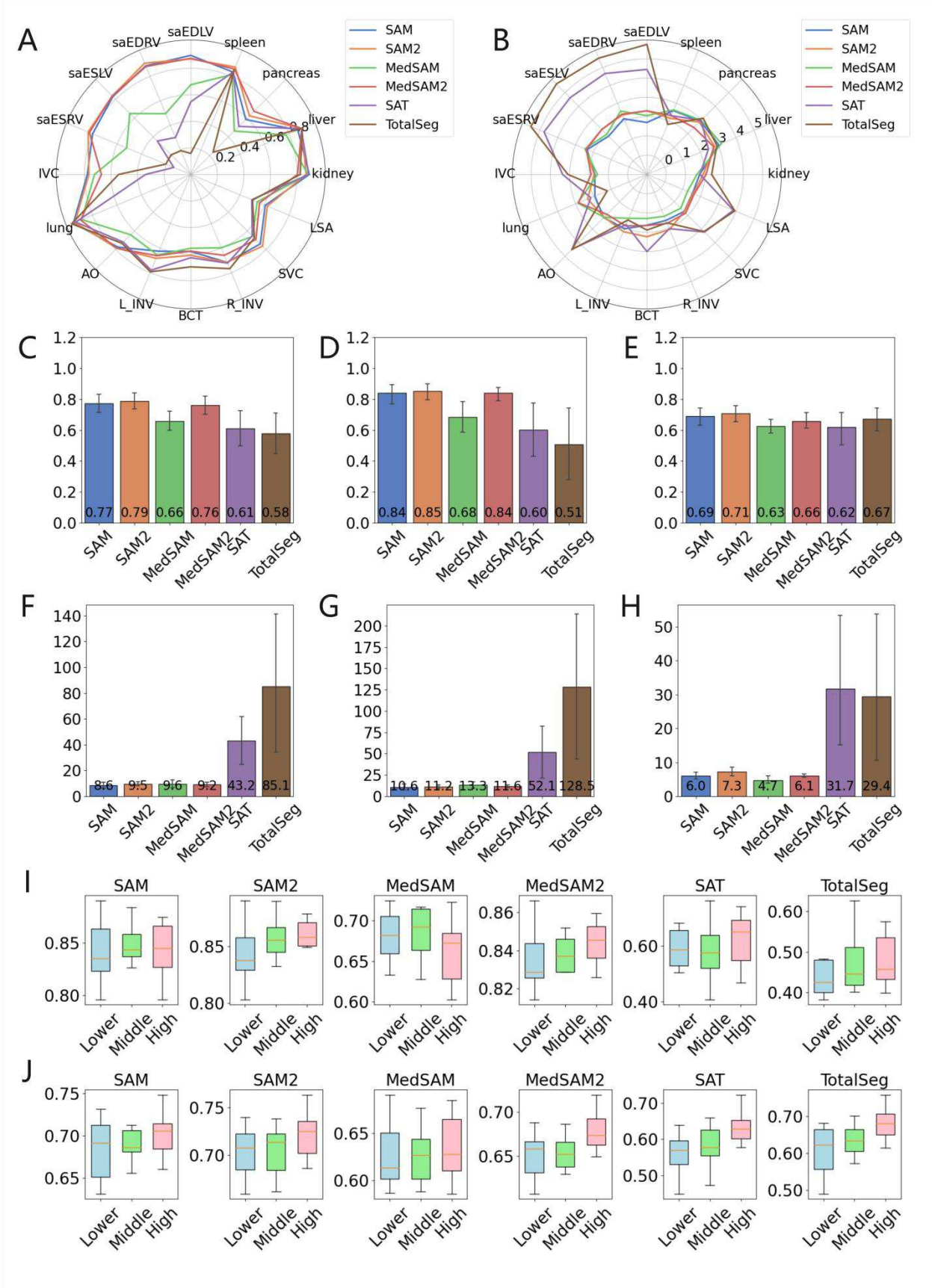

A
B
SAM
SAM2
MedSAM
MedSAM2
SAT
TotalSeg
saEDLV
spleen
pancreas
liver
kidney
LSA
SVC
R_INV
BCT
L_INV
AO
lung
IVC
saESRV
saESLV
saEDRV
C
D
E
0.77 0.79 0.66 0.76 0.61 0.58
0.84 0.85 0.68 0.84 0.60 0.51
0.69 0.71 0.63 0.66 0.62 0.67
F
G
H
8.6 9.5 9.6 9.2 43.2 85.1
10.6 11.2 13.3 11.6 52.1 128.5
6.0 7.3 4.7 6.1 31.7 29.4
I
J
Lower
Middle
High

Figure 2 | Segmentation accuracy is highly variable across models, anatomical targets, and task difficulty. A,**B**, Per-organ performance comparison based on **(A)** DSC and **(B)** HD95. Bbox-prompted models generally outperform others but struggle with vessels, while text-prompted and prompt-free models fail on complex cardiac structures. **C-H**, Overall accuracy summary for DSC **(C-E)** and HD95 **(F-H)** across all 16 organs, 7 large organs, and 9 vessels, illustrating the performance drop on intricate vascular structures. **I**,**J**, Model accuracy on a 30-case expert-annotated subset, stratified by manual annotation consistency (a proxy for task difficulty). Performance is shown for **(I)** large organs and **(J)** vessels across cases with low, medium, and high inter-rater agreement.

*EDLV, End-Diastolic Left Ventricle; ESLV, End-Systolic Left Ventricle; EDRV, End-Diastolic Right Ventricle; ESRV, End-Systolic Right Ventricle; IVC, Inferior Vena Cava; AO, Aorta; L_INV, left brachiocephalic vein; R_INV, right brachiocephalic vein; BCT, Brachiocephalic Trunk; SVC, Superior Vena Cava; LSA, Left Subclavian Artery.

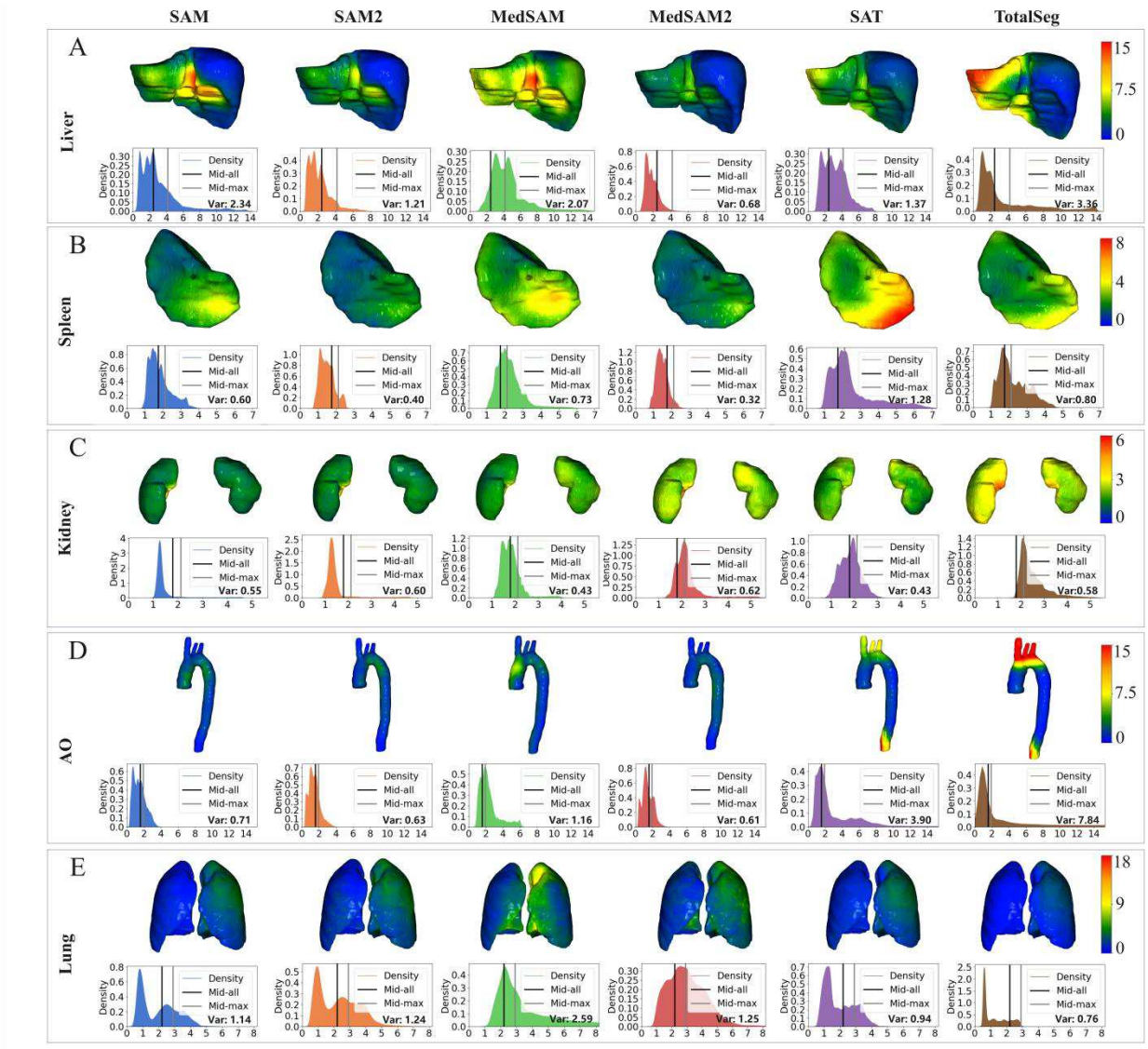


Figure 3 | Spatial segmentation error maps reveal distinct error topographies across SFM architectures. A-E, Population-averaged segmentation errors are visualized on 3D anatomical atlases for the liver **(A)**, spleen **(B)**, kidney **(C)**, aorta **(D)**, and lung **(E)**. Color intensity corresponds to the mean distance to the ground-truth segmentation, where red indicates regions of high error and blue indicates low error.

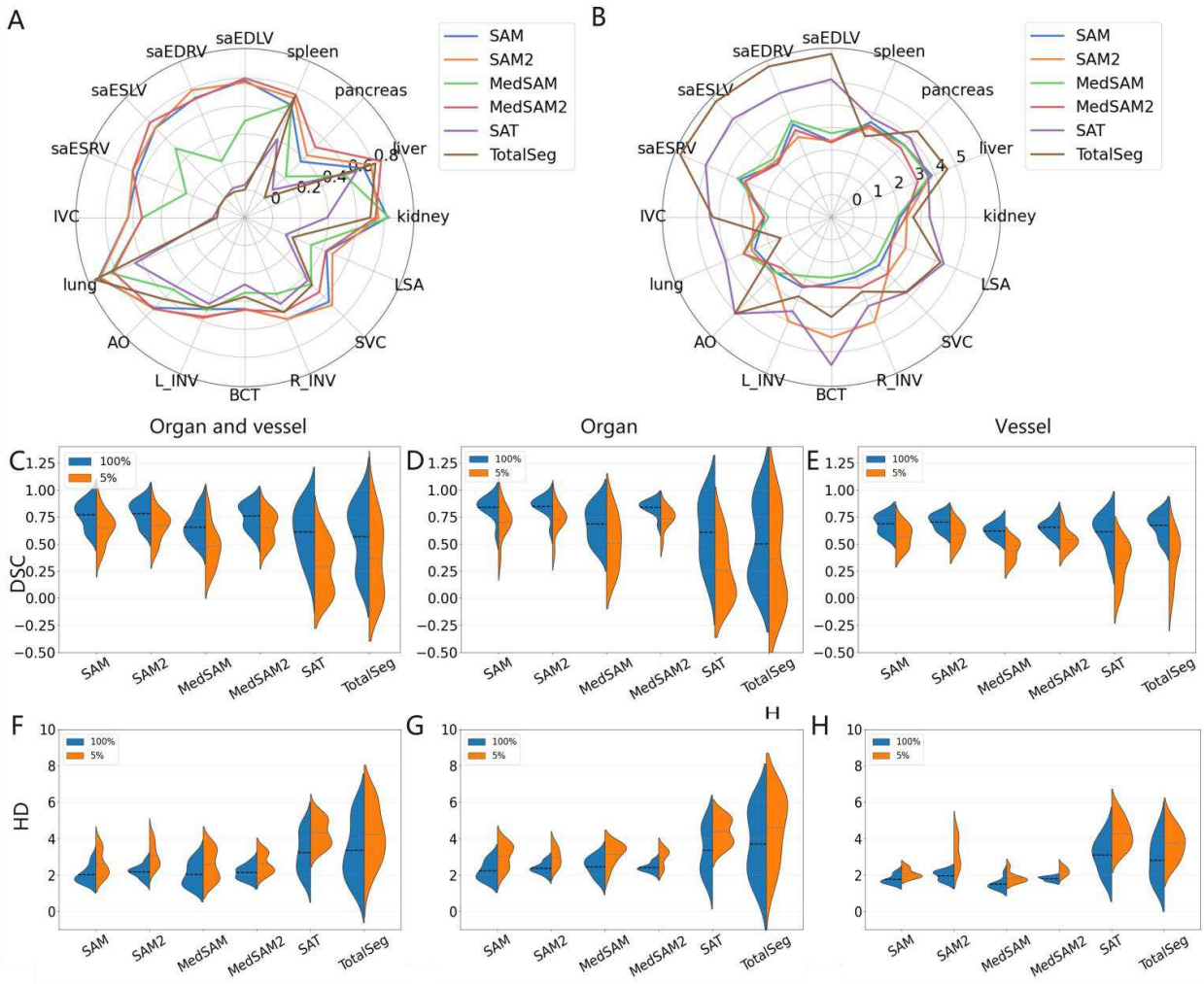


Figure 4 | Robustness test reveals the clinical brittleness of text-prompted and prompt-free SFMs. Model robustness was evaluated by assessing performance on the most challenging 5% of cases for each organ. **A,B,** Per-organ accuracy (DSC and HD95) on this "worst-case" cohort shows that while bbox-prompted models maintain reasonable performance, SAT and TotalSeg suffer a catastrophic performance collapse. **C-H,** Direct comparison of performance distributions between the worst 5% of cases (left violins) and all cases (right violins) for both DSC (**C-E**) and HD95 (**F-H).** The significant drop in median performance and wider distributions for SAT and TotalSeg in the worst-case cohort highlight their lack of robustness.

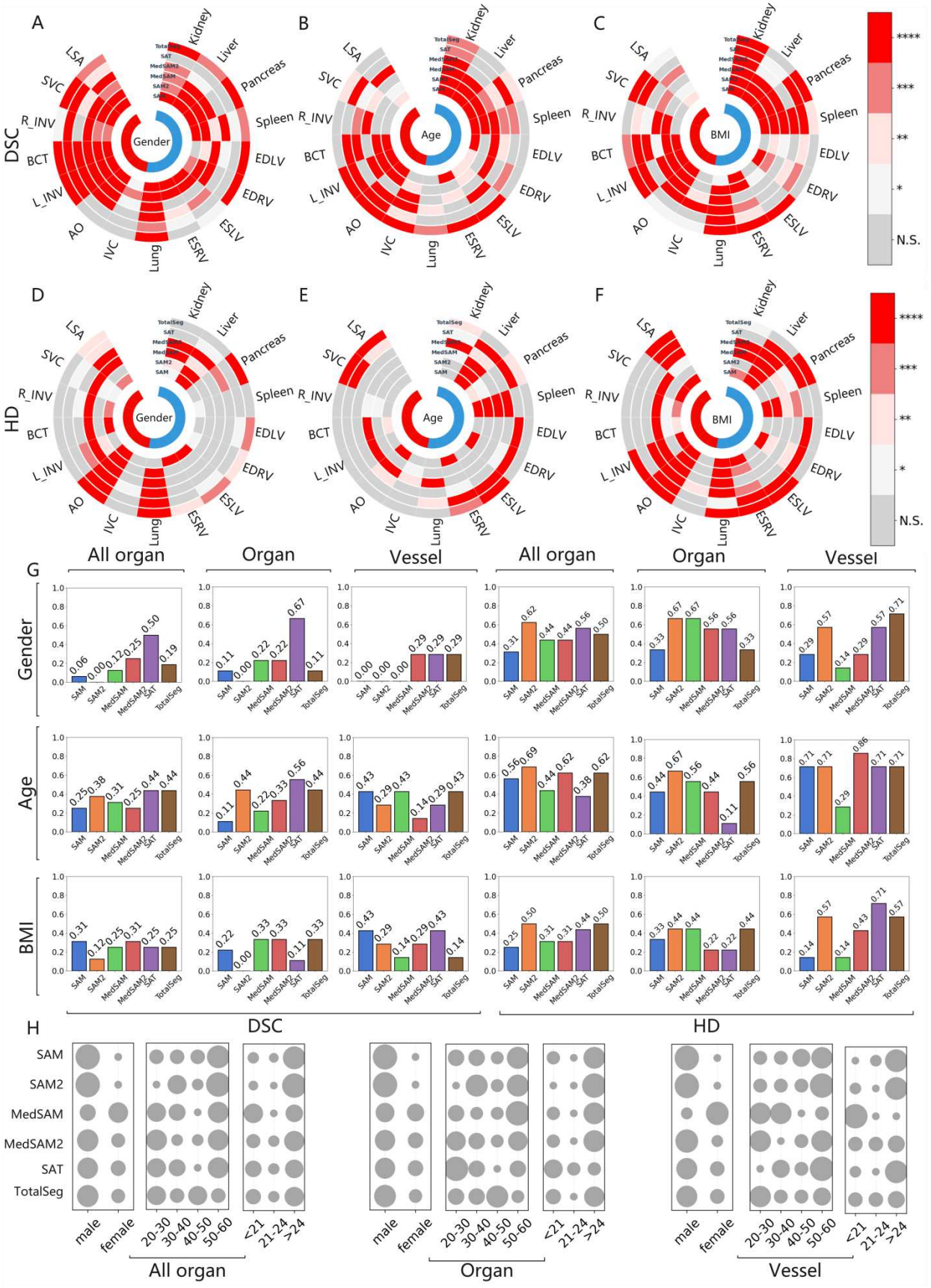

A
B
C
D
E
F
DSC
HD
Gender
Age
BMI
Kidney
Liver
Pancreas
Spleen
EDLV
EDRV
ESLV
ESRV
Lung
IVC
AO
L_INV
BCT
R_INV
SVC
LSA
****
***
**
*
N.S.
G
All organ
Organ
Vessel
SAM
SAM2
MedSAM
MedSAM2
SAT
TotalSeg
H
male
female
20-30
30-40
40-50
50-60
<21
21-24
>24

Figure 5 | Demographic fairness analysis reveals a critical accuracy-fairness trade-off. Model performance was evaluated for significant disparities across gender, age, and BMI subgroups. **A-F,** Statistical significance of performance bias for each of the 16 organs, measured by DSC (**A-C**) and HD95 (**D-F**). Each point represents an organ, with asterisks indicating a significant bias detected by one-way ANOVA. The results show that the most accurate models (e.g., SAM2) exhibit the most frequent and significant biases. **G,** Overall fairness score, defined as the proportion of organs with no significant performance bias ($p \geq 0.05$). This summary highlights that less accurate models like SAT are often fairer overall. **H,** A bias portrait visualizing the direction of systematic performance advantages. The size of each point corresponds to the number of organs for which a model performed best in a specific demographic subgroup, revealing, for example, a consistent male preference for most models.

Statistical significance: * $p<0.05$; ** $p<0.01$; *** $p<0.001$; **** $p<0.0001$; N.S., not significant.

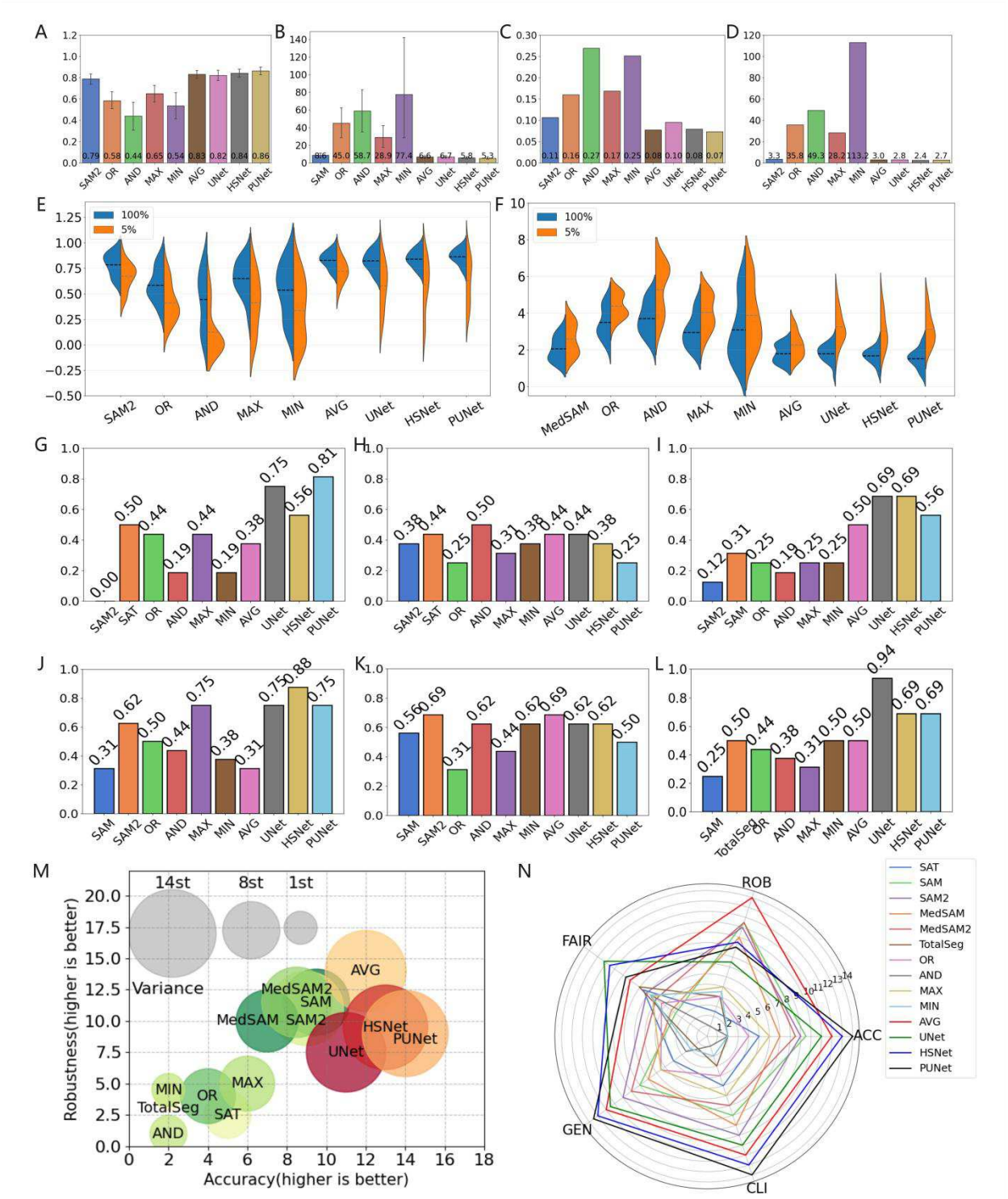


Figure 6 | Ensemble learning strategies resolves performance trade-offs and dominates individual models. Performance of ensemble strategies of five post-processing fusion and two multi-source distillation was compared against the best-performing individual models across all

evaluation dimensions. **A,B,** The AVG fusion and models of distillation achieve the highest overall accuracy (DSC and HD95). **C,D,** It also exhibits the lowest performance variance across organs, indicating superior generalization. **E,F,** For robustness, AVG fusion maintains the highest accuracy on the most challenging 5% of cases. **G-I,** Crucially, the AVG fusion and models of distillation maintain competitive fairness across gender, age, and BMI attributes, effectively resolving the accuracy-fairness trade-off. **M,N,** A holistic performance summary visually confirms that the ensemble learning strategy provides the best overall balance, dominating all other approaches across accuracy, robustness, fairness and generalization.

Table 1 | Clinical utility for downstream tasks reveals performance trade-offs resolved by ensemble learning. The table compares model performance on two clinical tasks: cardiac phenotype quantification (measured by absolute error %) and multi-organ volumetry (measured by relative volume error [VER] %). While individual models show conflicting strengths, the AVG fusion and 3 distilled models provide superior or highly competitive performance across both tasks.

	Baseline-SFMs						**Ensemble strategies**			
Model	SAM	SAM2	MedSAM	MedSAM2	SAT	TotalSeg	**AVG fusion**	**UNet**	**HSNet**	**PUNet**
LVEDV(mL)	24.15(0.18)	35.84(0.27)	59.78(0.44)	31.66(0.24)	41.28(0.31)	116.48(0.87)	15.00(0.11)	22.51(0.17)	6.22(0.05)	4.14(0.03)
LVESV(mL)	8.66(0.17)	11.08(0.22)	23.76(0.45)	15.08(0.30)	50.27(1.12)	41.90(0.80)	4.35(0.09)	5.70(0.12)	5.89(0.12)	3.32(0.07)
LVSV(mL)	19.68(0.26)	25.91(0.32)	36.02(0.47)	16.96(0.22)	45.68(0.60)	77.42(0.95)	16.41(0.21)	24.82(0.32)	8.35(0.10)	6.22(0.08)
LVEF(%)	5.41(0.11)	5.09(0.09)	6.62(0.17)	3.06(0.06)	36.62(0.64)	803.91(14.48)	7.01(0.12)	7.42(0.13)	4.64(0.08)	2.86(0.05)
LVCO(L/min)	1.21(0.26)	1.58(0.32)	2.18(0.47)	1.02(0.22)	2.77(0.60)	4.69(0.95)	1.00(0.21)	1.45(0.32)	0.48(0.10)	0.34(0.08)
RVEDV(mL)	24.49(0.18)	18.39(0.13)	47.48(0.33)	11.09(0.08)	171.40(1.33)	114.77(0.81)	19.48(0.13)	11.48(0.08)	8.04(0.06)	8.14(0.06)
RVESV(mL)	28.61(0.42)	21.01(0.31)	14.74(0.21)	14.19(0.21)	224.15(3.63)	54.58(0.78)	7.76(0.11)	10.52(0.15)	7.88(0.12)	7.04(0.10)
RVSV(mL)	12.54(0.23)	9.36(0.20)	34.73(0.52)	8.18(0.14)	64.47(0.96)	70.21(1.02)	14.71(0.21)	10.34(0.16)	9.47(0.14)	11.35(0.18)
RVEF(%)	10.64(0.25)	7.99(0.23)	14.37(0.33)	6.66(0.16)	46.42(0.95)	3180.29(63.36)	5.58(0.12)	6.34(0.15)	5.41(0.12)	5.70(0.14)
kidney(%)	0.10	0.14	0.05	0.3	0.05	0.24	0.06	0.06	0.07	0.08
liver(%)	0.18	0.11	0.11	0.1	0.05	0.14	0.06	0.04	0.05	0.02
pancreas(%)	0.5	0.51	0.26	0.56	0.30	0.53	0.35	0.3	0.33	0.18
spleen(%)	0.19	0.13	0.13	0.23	0.12	0.21	0.08	0.06	0.07	0.02
IVC(%)	0.25	0.33	0.11	0.49	0.45	0.39	0.13	0.22	0.20	0.05
lung(%)	0.06	0.09	0.12	0.05	0.05	0.04	0.05	0.19	0.20	0.19
AO(%)	0.32	0.32	0.24	0.32	0.18	0.25	0.15	0.16	0.23	0.06
L_INV(%)	0.53	0.45	0.11	0.49	0.16	0.15	0.33	0.28	0.38	0.13
BCT(%)	0.56	0.54	0.14	0.59	0.41	0.29	0.42	0.38	0.48	0.20
R_INV(%)	0.36	0.37	0.2	0.49	0.36	0.30	0.28	0.32	0.34	0.06
SVC(%)	0.35	0.33	0.28	0.43	0.27	0.26	0.18	0.26	0.23	0.07
LSA(%)	0.44	0.50	0.14	0.6	0.56	0.59	0.35	0.43	0.47	0.23

References

1. Sun, Y., Wang, L., Li, G., Lin, W. & Wang, L. A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. *Nat. Biomed. Eng* **9**, 521–538 (2025).
2. Wang, J. *et al.* Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nat Med* **31**, 609–617 (2025).
3. Kirillov, A. *et al.* Segment Anything. in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 3992–4003 (IEEE, Paris, France, 2023). doi:10.1109/ICCV51070.2023.00371.
4. Yan, S. *et al.* A multimodal vision foundation model for clinical dermatology. *Nat Med* **31**, 2691–2702 (2025).
5. Ma, J. & Wang, B. Towards foundation models of biological image segmentation. *Nat Methods* **20**, 953–955 (2023).
6. Ma, J. *et al.* Segment Anything in Medical Images and Videos: Benchmark and Deployment. Preprint at https://doi.org/10.48550/arXiv.2408.03322 (2024).
7. Huang, Y. *et al.* Segment anything model for medical images? *Medical Image Analysis* **92**, 103061 (2024).
8. Mattjie, C. *et al.* Zero-Shot Performance of the Segment Anything Model (SAM) in 2D Medical Imaging: A Comprehensive Evaluation and Practical Guidelines. in *2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)* 108–112 (2023). doi:10.1109/BIBE60311.2023.00025.
9. Shi, P. *et al.* Generalist Vision Foundation Models for Medical Imaging: A Case Study of Segment Anything Model on Zero-Shot Medical Segmentation. *Diagnostics* **13**, 1947 (2023).
10. Pai, S. *et al.* Foundation model for cancer imaging biomarkers. *Nat Mach Intell* **6**, 354–367 (2024).
11. Archit, A. *et al.* Segment Anything for Microscopy. *Nat Methods* **22**, 579–591 (2025).
12. Huai, Z., Tang, H., Li, Y., Chen, Z. & Li, X. Leveraging Segment Anything Model for Source-Free Domain Adaptation via Dual Feature Guided Auto-Prompting. *IEEE Transactions on Medical Imaging* 1–1 (2025) doi:10.1109/TMI.2025.3587733.
13. Zhang, C. *et al.* A Survey on Segment Anything Model (SAM): Vision Foundation Model Meets Prompt Engineering. Preprint at https://doi.org/10.48550/arXiv.2306.06211 (2024).
14. He, S. *et al.* Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets. Preprint at https://doi.org/10.48550/arXiv.2304.09324 (2023).
15. Mazurowski, M. A. *et al.* Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis* **89**, 102918 (2023).
16. Ma, J. *et al.* Segment anything in medical images. *Nat Commun* **15**, 654 (2024).
17. Zhao, Z. *et al.* Large-Vocabulary Segmentation for Medical Images with Text Prompts. Preprint at https://doi.org/10.48550/arXiv.2312.17183 (2025).
18. Wasserthal, J. *et al.* TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence* (2023) doi:10.1148/ryai.230024.

19. Ricci Lara, M. A., Echeveste, R. & Ferrante, E. Addressing fairness in artificial intelligence for medical imaging. *Nat Commun* **13**, 4581 (2022).
20. Chen, R. J. *et al.* Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng* **7**, 719–742 (2023).
21. Mhasawade, V., Zhao, Y. & Chunara, R. Machine learning and algorithmic fairness in public and population health. *Nat Mach Intell* **3**, 659–666 (2021).
22. Li, Q. *et al.* An Empirical Study on the Fairness of Foundation Models for Multi-Organ Image Segmentation. Preprint at https://doi.org/10.48550/arXiv.2406.12646 (2024).
23. Striving for health equity with machine learning. *Nat Mach Intell* **3**, 653–653 (2021).
24. Abràmoff, M. D. *et al.* Considerations for addressing bias in artificial intelligence for health equity. *npj Digit. Med.* **6**, 170 (2023).
25. Huang, M. *et al.* A Clinical Decision Support Framework for Heterogeneous Data Sources. *IEEE Journal of Biomedical and Health Informatics* **22**, 1824–1833 (2018).
26. Kumar, G. *et al.* Data Harmonization for Heterogeneous Datasets: A Systematic Literature Review. *Applied Sciences* **11**, 8275 (2021).
27. Ma, J. *et al.* PathBench: A comprehensive comparison benchmark for pathology foundation models towards precision oncology. Preprint at https://doi.org/10.48550/arXiv.2505.20202 (2025).
28. Maier-Hein, L. *et al.* Metrics reloaded: recommendations for image analysis validation. *Nat Methods* **21**, 195–212 (2024).
29. Ravi, N. *et al.* SAM 2: Segment Anything in Images and Videos. Preprint at https://doi.org/10.48550/arXiv.2408.00714 (2024).
30. Zhu, J., Hamdi, A., Qi, Y., Jin, Y. & Wu, J. Medical SAM 2: Segment medical images as video via Segment Anything Model 2. Preprint at https://doi.org/10.48550/arXiv.2408.00874 (2024).
31. Zhao, Z. *et al.* One Model to Rule them All: Towards Universal Segmentation for Medical Images with Text Prompts. Preprint at https://doi.org/10.48550/arXiv.2312.17183 (2024).
32. Balendran, A. *et al.* A scoping review of robustness concepts for machine learning in healthcare. *npj Digit. Med.* **8**, 38 (2025).
33. Xu, Z. *et al.* Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *npj Digit. Med.* **7**, 286 (2024).
34. Xu, Z., Li, J., Yao, Q., Li, H. & Zhou, S. K. Fairness in Medical Image Analysis and Healthcare: A Literature Survey. Preprint at https://doi.org/10.36227/techrxiv.24324979.v1 (2023).
35. Colunga, A. L. *et al.* Deep phenotyping of cardiac function in heart transplant patients using cardiovascular system models. *The Journal of Physiology* **598**, 3203–3222 (2020).
36. Bevilacqua, M. U. *et al.* CT of Kidney Volume in Autosomal Dominant Polycystic Kidney Disease: Accuracy, Reproducibility, and Radiation Dose. *Radiology* (2019) doi:10.1148/radiol.2019181830.

37. Grote, T., Freiesleben, T. & Berens, P. Foundation models in healthcare require rethinking reliability. *Nat Mach Intell* **6**, 1421–1423 (2024).
38. Zheng, H. *et al.* Learning from models beyond fine-tuning. *Nat Mach Intell* **7**, 6–17 (2025).
39. Zhang, K. *et al.* A generalist vision–language foundation model for diverse biomedical tasks. *Nat Med* **30**, 3129–3141 (2024).
40. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
41. Tu, T. *et al.* Towards Generalist Biomedical AI. *NEJM AI* (2024) doi:10.1056/AIoa2300138.
42. Wang, G. *et al.* SAM-Med3D-MoE: Towards a Non-Forgetting Segment Anything Model via Mixture of Experts for 3D Medical Image Segmentation. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* (eds. Linguraru, M. G. et al.) 552–561 (Springer Nature Switzerland, Cham, 2024). doi:10.1007/978-3-031-72114-4_53.
43. Li, Y., Zhang, L., Liang, Y. & Xie, P. AM-SAM: Automated Prompting and Mask Calibration for Segment Anything Model. Preprint at https://doi.org/10.48550/arXiv.2410.09714 (2024).
44. Wang, C. *et al.* Protocol for Brain Magnetic Resonance Imaging and Extraction of Imaging-Derived Phenotypes from the China Phenobank Project. *Phenomics* **3**, 642–656 (2023).
45. Wang, C. *et al.* A Protocol for Body MRI/CT and Extraction of Imaging-Derived Phenotypes (IDPs) from the China Phenobank Project. *Phenomics* **4**, 594–616 (2024).
46. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds. Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) vol. 9351 234–241 (Springer International Publishing, Cham, 2015).
47. Zhang, W. *et al.* HSNet: A hybrid semantic network for polyp segmentation. *Computers in Biology and Medicine* **150**, 106173 (2022).
48. Wang, W. *et al.* Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions. in 568–578 (2021).
49. Rezatofighi, H. *et al.* Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. in 658–666 (2019).
50. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423 (1948).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx