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: Bridging the Translational Gap for Medical Segmentation Foundation Models Through Principled Model Synergy

Qing Li
Fudan University
Yizhe Zhang
Nanjing University of Science and Technology
Xin Guo
Shanghai Academy of Artificial Intelligence for Science
Haosen Zhang
Fudan University
Mo Yang
Fudan University
Mengting Sun
Fudan University
Longyu Sun
Fudan University
Haoyang Zhang
Fudan University
Junhong Liu
Fudan University
Yan Li
Shanghai Jiaotong University
Yajing Zhang
GE Healthcare
Shuo Wang
Fudan University
Chengyan Wang
wangcy@fudan.edu.cn
Fudan University

Article

Keywords: Foundation models, Medical image segmentation, Benchmark evaluation, Algorithmic fairness, Model synergy

Posted Date: June 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9781531/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

OrganSegBench: Bridging the Translational Gap for Medical Segmentation Foundation Models Through Principled Model Synergy

Qing Li, Yizhe Zhang, Xin Guo, Haosen Zhang, Yan Li, Mo Yang, Yajing Zhang, Mengting Sun, Longyu Sun, Haoyang Zhang, Junhong Liu, Shuo Wang, Chengyan Wang

***Abstract*—The clinical translation of Segmentation Foundation Models (SFMs) is currently impeded by a critical misalignment between general-purpose artificial intelligence (AI) capabilities and the stringent requirements of healthcare, particularly regarding robustness, fairness and clinical performance. While SFMs show promise in computer vision, existing benchmarks often provide overoptimistic performance estimates due to potential training data leakage and a narrow focus on accuracy metrics. To address this translational gap, we introduce "OrganSegBench", a multi-dimensional evaluation framework grounded in two high-quality, independent cohorts: the in-house China PhenoBank (CPB) (N = 701; 16 organs) and the UK Biobank (UKB) (N = 792; 9 organs), with detailed demographic metadata. Through a systematic assessment of six state-of-the-art SFMs across five clinical dimensions, i.e., accuracy, generalization, robustness, fairness, and clinical utility, we demonstrate that no single monolithic model achieves optimal performance across all tasks. Our analysis reveals a persistent trade-off between segmentation accuracy and demographic fairness. To resolve these limitations, we propose a paradigm shift from model scale to model synergy, implementing two ensemble strategies: training-free fusion and multi-source knowledge distillation. Notably, these synergistic approaches decisively outperformed every individual SFM across all dimensions, effectively resolving the accuracy-fairness dilemma and enhancing cross-cohort stability. These findings expose the inherent brittleness of current monolithic SFMs and establish principled model synergy as a superior, practical pathway toward building safe, equitable, and clinically trustworthy AI.**

***Index Terms*—Foundation models, Medical image segmentation, Benchmark evaluation, Algorithmic fairness, Model synergy**

In this paper, the co-corresponding authors are Yizhe Zhang, Shuo Wang and Chengyan Wang, who made substantial contributions to the research and its direction.

Qing Li, Haosen Zhang, Mo Yang, Mengting Sun, Longyu Sun, Haoyang Zhang, Junhong Liu are with the Human Phenome Institute, Fudan University, Shanghai, China. (e-mail: 13188816530@163.com; haosen.zhang24@imperial.ac.uk; myang5061@outlook.com; sunmengting150@163.com; 2501111769@stu.pku.edu.cn; 25212030015@m.fudan.edu.cn; 2609668279@qq.com).

Yizhe Zhang is with School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, China(e-mail: zhangyizhe@njust.edu.cn).

Xin Guo is with Shanghai Academy of Artificial Intelligence for Science, Shanghai, China (e-mail: guoxin@sais.com.cn).

Yan Li is with the Department of Radiology, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China (e-mail: ly40730@rjh.com.cn).

Yajing Zhang is with the GE Healthcare, Beijing, China (e-mail: yajing.zhang@gehealthcare.com).

Shuo Wang is with the Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, Shanghai, China (e-mail: shuowang@fudan.edu.cn).

Chengyan Wang is with Shanghai Pudong Hospital and Human Phenome Institute, Fudan University, Shanghai, China (e-mail: wangcy@fudan.edu.cn).

I. Introduction

Medical foundation models have emerged as a transformative force in healthcare [1], [2], particularly in medical imaging where they offer unprecedented potential for comprehensive, non-invasive disease assessment. At the forefront of this evolution is the pursuit of fully automated multi-organ and lesion segmentation, a task that demands foundation models with high precision and adaptability. Segmentation Foundation Models (SFMs), exemplified by the Segment Anything Model (SAM), have fundamentally reshaped the landscape of computer vision [3]. By learning powerful, transferable feature representations from vast, heterogeneous datasets, these models exhibit remarkable zero-shot generalization capabilities, even on unseen datasets within familiar tasks [4]–[7]. This paradigm shift promises to democratize clinical artificial intelligence (AI) by reducing the reliance on costly expert annotations and accelerating the development of adaptable medical systems [8], [9]. Their prompt-driven design, utilizing bounding boxes (bboxes) or text prompts, offers a flexible and intuitive interface for complex specialized tasks [10], [11].

However, the clinical translation of SFMs into the high-stakes medical environments remains a formidable challenge [12], [13]. The inherent complexity of clinical data, characterized by subtle inter-patient variability and distinct modality-specific properties, often leads generalist models to exhibit suboptimal and unstable performance [7]. While domain-specific adaptations such as MedSAM [14], or advanced 3D architectures like SAT [15] and TotalSegmentator [16], have improved segmentation accuracy, they have simultaneously exposed deeper, unresolved issues regarding algorithmic fairness [17], [18]. Significant performance biases across demographic subgroups persist, posing a critical barrier to safe and equitable deployment in clinical practice, where health equity is a paramount concern [19].

This translational gap is exacerbated by the limitations of current benchmarking practices. Most evaluations rely exclusively on public datasets, which are prone to "data leakage", the inadvertent inclusion of test data in a model's pre-training corpus, resulting in artificially inflated performance estimates [20]. Furthermore, existing benchmarks remain disproportionately focused on technical metrics like the Dice Similarity

Coefficient (DSC), while neglecting clinically critical dimensions such as robustness under challenging conditions and demographic equity across diverse populations [21]–[23]. This misalignment between benchmark performance and real-world utility leads to an overoptimistic assessment of clinical readiness, further widening the gap between laboratory research and bedside application.

To address these critical constraints, we introduce "OrganSegBench", a comprehensive framework designed bridge the gap between algorithmic research and practical clinical applications (Fig. 1). Our work is anchored in two high-quality, independent resources: China PhenoBank (CPB) (N = 701; 16 organs, see Fig S1 and Table S1) and UK Biobank (UKB) (N = 792; 9 organs). Specifically, by utilizing CPB, a dataset released for the first time by our institution via PhenoBank (*https://www.phenobank.org.cn/*), we eliminate the risk of data leakage and ensure rigorous quality control. Crucially, while the raw MRI scans of the UKB are accessible via open institutional application, the public repository entirely lacks localized multi-organ or vessel segmentation ground truths; to establish an unassailable evaluation yardstick, we invited board-certified radiologists to perform meticulous, slice-by-slice manual delineation across all 792 external subjects. Leveraging these dual cohorts, we conducted a systematic, five-dimensional evaluation of six prominent SFMs, assessing their accuracy, generalization, robustness, fairness, and clinical utility. We have conducted strict quality control on the annotations of both CPB and UKB to guarantee the consistency of label quality across the two datasets. Additionally, we fully leveraged the demographic characteristics of both datasets to conduct in-depth fairness analyses. Using this OrganSegBench framework, we conducted a multi-dimensional evaluation of six prominent SFMs, assessing their accuracy, generalization, robustness, demographic fairness, and clinical utility.

Crucially, our study challenges the dominant "bigger is better" trajectory of AI development. To enhance the performance and cost-effectiveness of clinical deployment, we propose two ensemble strategies: **training-free fusion** and **multi-source knowledge distillation**. This work, which extends our preliminary findings [24], demonstrates that while individual monolithic SFMs exhibit inherent brittleness, synergistic ensemble approaches consistently deliver superior performance across all evaluation dimensions. This work establishes a new benchmark and a superior deployment paradigm, moving the field toward a more safe, equitable, and synergistic generation of medical AI.

II. Result

A. Overview of OrganSegBench and its multi-dimensional evaluation

As illustrated in Fig.1 A-F, OrganSegBench provides a standardized framework designed to evaluate diverse SFMs under high-stakes clinical constraints. Its adaptable architecture accommodates heterogeneous SFM paradigms, including bbox-prompted, text-prompted, and prompt-free approaches, thereby enabling unified benchmarking across imaging modalities and providing a modular foundation for future clinical deployment platforms. The framework rigorously assesses SFM performance across five critical dimensions: accuracy, generalization, robustness, fairness, and clinical utility.

Initial comparative analysis revealed a clear performance hierarchy: bbox-prompted SFMs (SAM [3], MedSAM [14], SAM2 [25], MedSAM2 [26]) consistently outperformed text-prompted (SAT [15]) and prompt-free (TotalSeg [16]) models 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.1 H). To probe the spatial characteristics of 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 while guiding 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 independent SFMs with simple logical operations. The second strategy proposes multi-source knowledge distillation to train a single, specialized model from their combined knowledge. Both paradigms decisively surpassed any individual monolithic model, demonstrating a lightweight yet effective pathway toward clinically robust and equitable AI.

B. Profound performance heterogeneity and cross-cohort reversals

As illustrated in Fig. 2, our multi-cohort investigation reveals profound performance heterogeneity among state-of-the-art SFMs. A baseline hierarchy emerged across both cohorts: bbox-prompted architectures generally maintained higher DSC than SAT or TotalSeg. However, a critical performance reversal emerged for the medically fine-tuned MedSAM. In the CPB cohort, the generalist SAM outperformed MedSAM (mean DSC 0.77 vs. 0.66), aligning with the assumption that domain adaptation can sometimes erode generalizability. Strikingly, this trend completely reversed in the independent UKB cohort, where MedSAM achieved the highest overall DSC (0.61) among all six models, surpassing SAM (0.59), SAM2 (0.59), and MedSAM2 (0.55). This cross-cohort reversal directly challenges the notion of a "universal" performance ranking, highlighting that the effectiveness of domain-specific fine-tuning is contingent upon the distributional alignment between the target dataset and the model's pre-training corpus.

Furthermore, accuracy was heavily tied to anatomical complexity (Fig. 2A, B, D, E, G, H). While SFMs performed strongly on well-demarcated organs (e.g., liver, spleen), accuracy degraded substantially on complex targets like the pancreas and fine-grained vessels. Most SFMs exhibited a consistent alignment between task difficulty and accuracy, delivering superior performance on "basic" cases compared to "difficult" ones (Fig. 2I, J). These findings were further validated using 95% Hausdorff Distance (HD95) (Supplementary Fig. S2).

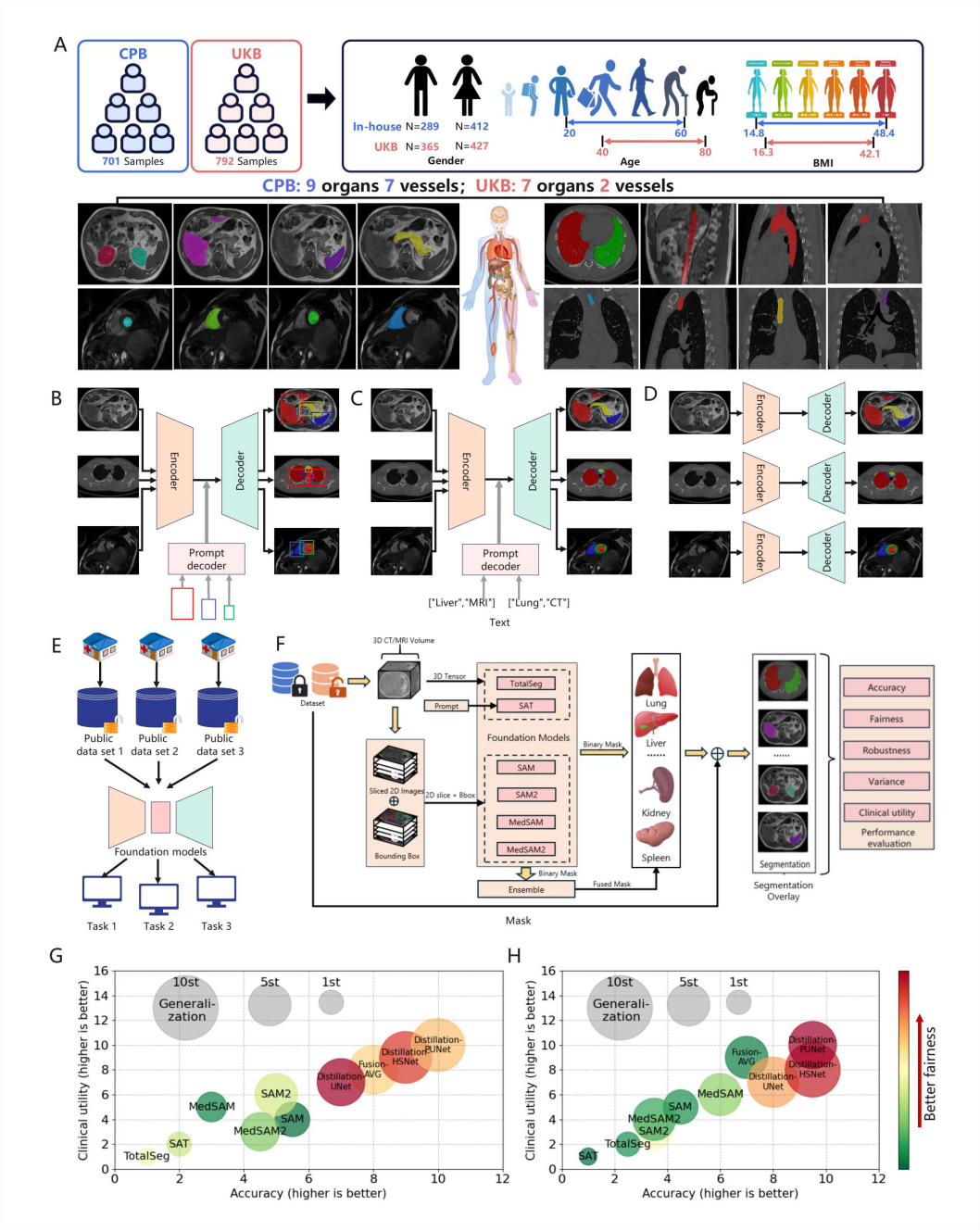


Fig. 1: **The OrganSegBench framework for multi-dimensional evaluation and synergistic enhancement of medical segmentation foundation models. A,** Data Resources and Demographic Diversity: Overview of the dual-cohort design, featuring the in-house China PhenoBank (CPB) (N = 701; 16 organs/vessels) and the UK Biobank (UKB) (N = 792; 9 organs/vessels), with detailed demographic profiles including gender, age, and BMI utilized for comprehensive fairness auditing. **B–D,** Evaluated SFM Architectural Paradigms: The benchmark encompasses diverse model architectures, including **(B)** bbox-prompted (e.g., SAM, MedSAM), **(C)** text-prompted (e.g., SAT), and **(D)** prompt-free (e.g., TotalSegmentator) approaches. **E–F,** Benchmark Workflow and Evaluation Pipeline: SFMs pre-trained on large-scale datasets **(E)** are subjected to a rigorous five-dimensional evaluation pipeline **(F)** assessing accuracy, generalization, robustness, demographic fairness, and clinical utility. **G–H,** Multi-dimensional Performance and Model Synergy: Summary of benchmark results on CPB **(G)** and UKB **(H)**, illustrating the performance trade-offs inherent in individual monolithic models and demonstrating the consistent superiority of the AVG fusion strategy and multi-source knowledge distillation in achieving a balanced, high-performance profile across all evaluation dimensions.

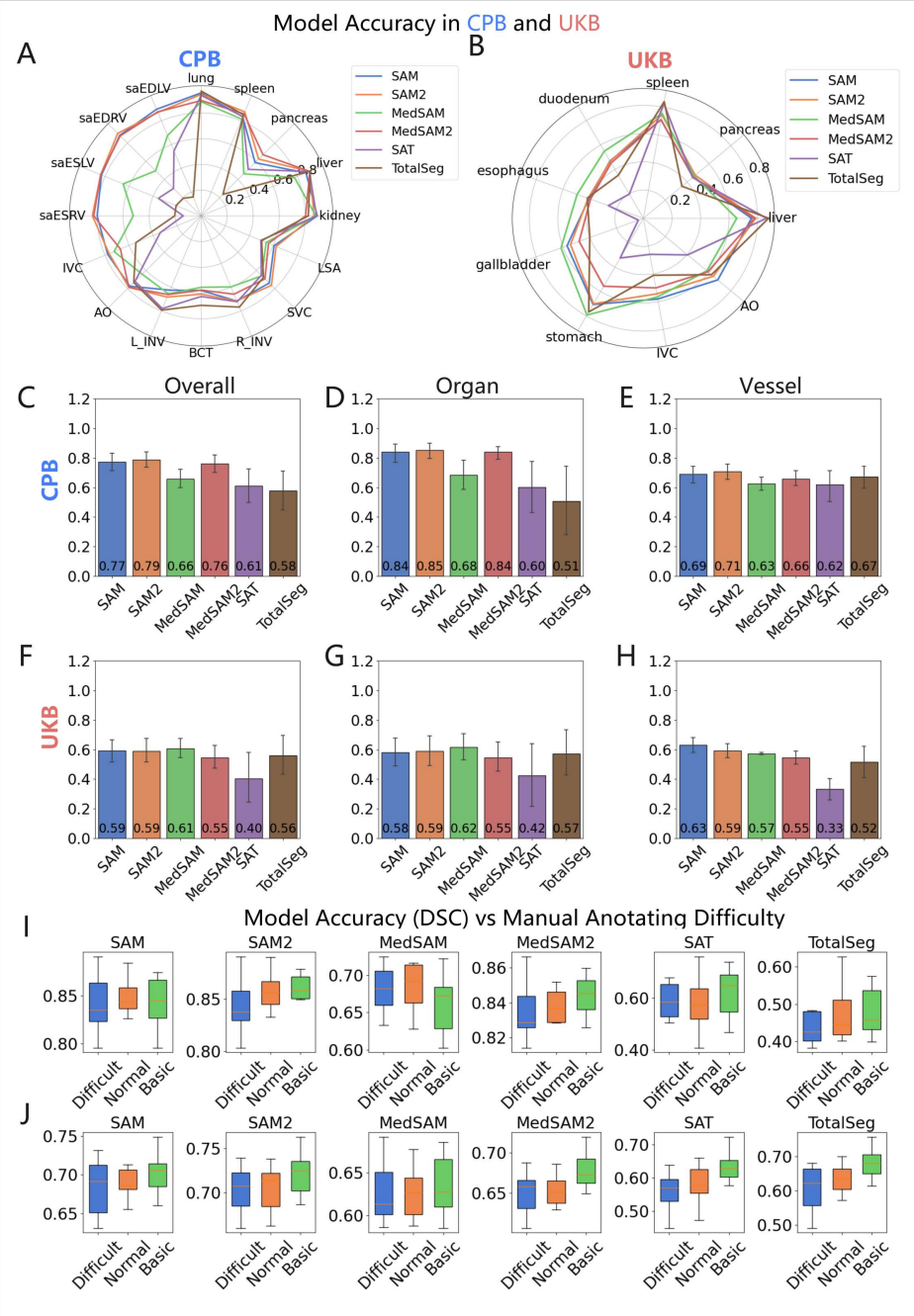


Fig. 2: **Quantitative assessment of segmentation accuracy across diverse anatomical structures and task difficulties. A, B,** Per-organ performance profiles: Radar plots illustrating the DSC for individual anatomical structures in the **(A)** CPB (in-house) and **(B)** UKB cohorts. Bbox-prompted models demonstrate superior accuracy on large organs but exhibit performance degradation on intricate vascular structures. **C–H,** Aggregated accuracy summary: Mean DSC scores across **(C, F)** all anatomical targets, **(D, G)** large organs, and (E, H) vessels for both CPB and UKB cohorts, highlighting the distinct performance hierarchy and the consistency of accuracy drops on vascular tasks. **I, J,** Task difficulty stratification: Model performance on a 30-case expert-annotated subset of CPB, categorized by manual annotation consistency (Basic, Normal, Complex) for **(I)** large organs and **(J)** vessels. Most SFMs exhibited a consistent alignment between task difficulty and accuracy, delivering superior performance on ”basic” cases compared to ”complex” ones. *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.

C. Spatial error heterogeneity beyond global metrics

To dissect SFM performance beyond global scores, we visualized and quantified spatial error distributions across seven representative anatomical structures (Fig. 3A–G). This analysis revealed stable, paradigm-specific patterns: bbox-prompted models predominantly exhibited errors in the core regions of organs while maintaining sharp boundaries. In contrast, SAT and TotalSeg were prone to catastrophic boundary errors, particularly on irregularly shaped organs like the duodenum and stomach.

Crucially, we observed that bbox-prompted models (especially SAM and SAM2) exhibited lower variance in error distribution (e.g., Var = 0.71 for CPB aorta), indicating more predictable performance. In contrast, SAT and TotalSeg showed significantly higher volatility (e.g., Var = 11.4 for UKB esophagus). Notably, even for anatomical structures where models achieved high aggregate DSC, such as the liver, we observed pronounced spatial heterogeneity in segmentation errors. These findings underscore that a single accuracy score masks critical error heterogeneity, necessitating spatial visualization to identify regions requiring expert oversight prior to clinical deployment.

D. Robustness testing reveals clinical brittleness in challenging cases

A model's true clinical viability is determined by its robustness in challenging, real-world cases [27]. We stress-tested each model on the most difficult 5% of cases for each organ, which were selected based on the bottom 5% values of segmentation performance metrics including DSC and 95% Hausdorff Distance (HD95). As shown in Fig. 4, a clear robustness gap emerged: while bbox-prompted models maintained stable quality, SAT and TotalSeg suffered catastrophic performance collapse on complex targets. The HD95 results, provided in the Supplementary Fig. S3, further validated these trends, showing explosive increases in boundary error for SAT and TotalSeg on challenging cases.

This failure mode is entirely invisible in average DSC scores. Even models that appear usable on routine samples can exhibit catastrophic unreliability in "worst-case" scenarios, posing severe risks for high-stakes clinical decision-making. Our cross-cohort findings confirm that current monolithic models are inherently brittle when faced with atypical anatomy or poor scan quality.

E. Pervasive demographic disparities and the accuracy–fairness trade-off

Beyond accuracy and robustness, the ethical deployment of medical AI hinges on algorithmic fairness. As reflected in the fairness heatmaps (Fig. 5A–F), all investigated models exhibited pervasive and statistically significant subgroup biases across gender, age, and BMI. Our auditing reveals that these biases are not isolated incidents but represent systematic performance inequities across nearly all anatomical structures in both independent cohorts.

We identified a complex accuracy–fairness trade-off: in the CPB cohort, high-accuracy monolithic models (SAM, SAM2) consistently exhibited the most frequent significant biases, while lower-accuracy models like SAT achieved more equitable performance across demographic strata (Fig. 5G, H). While this gender-based trade-off remained stable across both cohorts, the relationship proved to be highly dataset-dependent. For age and BMI attributes in the UKB dataset, the inverse accuracy–fairness relationship significantly diminished; SAT lost its unique fairness advantage and exhibited performance disparities comparable to high-accuracy models (Fig. 5G, H).

Furthermore, the "bias portrait" in Fig. 5H uncovers a systematic subgroup preference, with most models, particularly the SAM family, demonstrating markedly better performance on male, older, and higher BMI subjects in the CPB setting. These findings underscore that fairness is not an inherent model property but a dynamic outcome of model-data interaction. Consequently, clinical translation must move beyond the simplistic assumption that lower-accuracy models are inherently fairer, emphasizing instead the necessity for attribute-specific and cross-cohort fairness auditing prior to deployment and our results revealed that no single model achieves equitable performance across all demographic attributes and cohorts..

Notably, we only present DSC-based fairness results for both cohorts in the main figure, while all HD95-based fairness outcomes of in-house and UKB data are relegated to the Appendix. Supplementary Fig. S4 revealed that model fairness was metric-dependent: HD95 yielded better boundary segmentation equity than DSC in most cases.

F. Contradictory performance profiles and the clinical fragmentation of monolithic models

To bridge the gap between segmentation metrics and practical clinical translation, we evaluated the downstream utility of each standalone SFM across two routine clinical workflows: cardiac function analysis and multi-organ volumetry [28], [29]. This cross-cohort evaluation crystallized a fundamental clinical bottleneck: no single monolithic model offers a universally reliable solution, owing to profoundly contradictory performance profiles that vary unpredictably across clinical tasks, anatomical organs, and population.

For the quantification of cardiac functional phenotypes (e.g., ejection fraction, cardiac output), the high segmentation fidelity of bbox-prompted architectures translated directly into superior downstream clinical utility (Table I). However, this established performance hierarchy fundamentally inverted when transitioning to multi-organ volumetry for non-cardiac structures. In this abdominal domain, text-prompted (SAT) and prompt-free (TotalSeg) models decisively outperformed the otherwise highly accurate SAM-family models. The competent anatomical boundary delineation achieved by SAT and TotalSeg on large abdominal organs rendered them the more effective choice for generalized volumetry. This paradox exposes an untenable clinical dilemma: clinicians are forced to choose between a model optimized for localized cardiac phenotyping and another suited for holistic abdominal assessment, rendering a unified, streamlined clinical workflow impossible under any single monolithic SFM architecture.

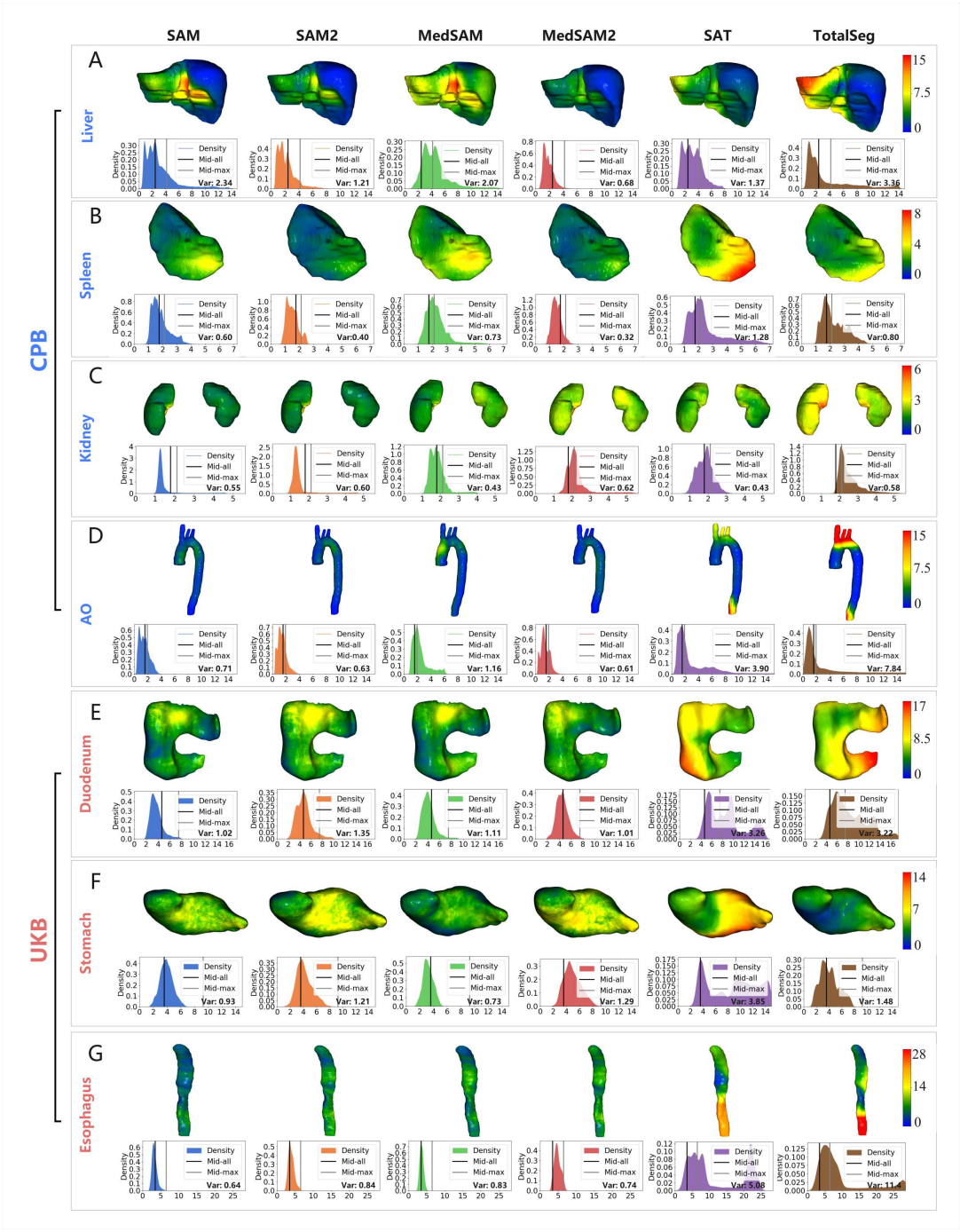


Fig. 3: **Spatial distribution of segmentation errors revealing paradigm-specific error topographies across SFM architectures** Population-averaged segmentation errors are projected onto 3D anatomical atlases for the **(A)** liver, **(B)** spleen, **(C)** kidney, and **(D)** aorta from the CPB cohort, and the **(E)** duodenum, **(F)** stomach, and **(G)** esophagus from the UKB cohort. Color intensity represents the mean Euclidean distance to the ground-truth segmentation, where red gradients denote regions of high spatial error and blue denotes high fidelity. Notably, even for structures with high aggregate DSC (e.g., liver), pronounced spatial heterogeneity is observed. Bbox-prompted models predominantly exhibit interior errors while preserving boundary precision, whereas text-prompted and prompt-free models are characterized by catastrophic boundary failures across diverse anatomical targets.

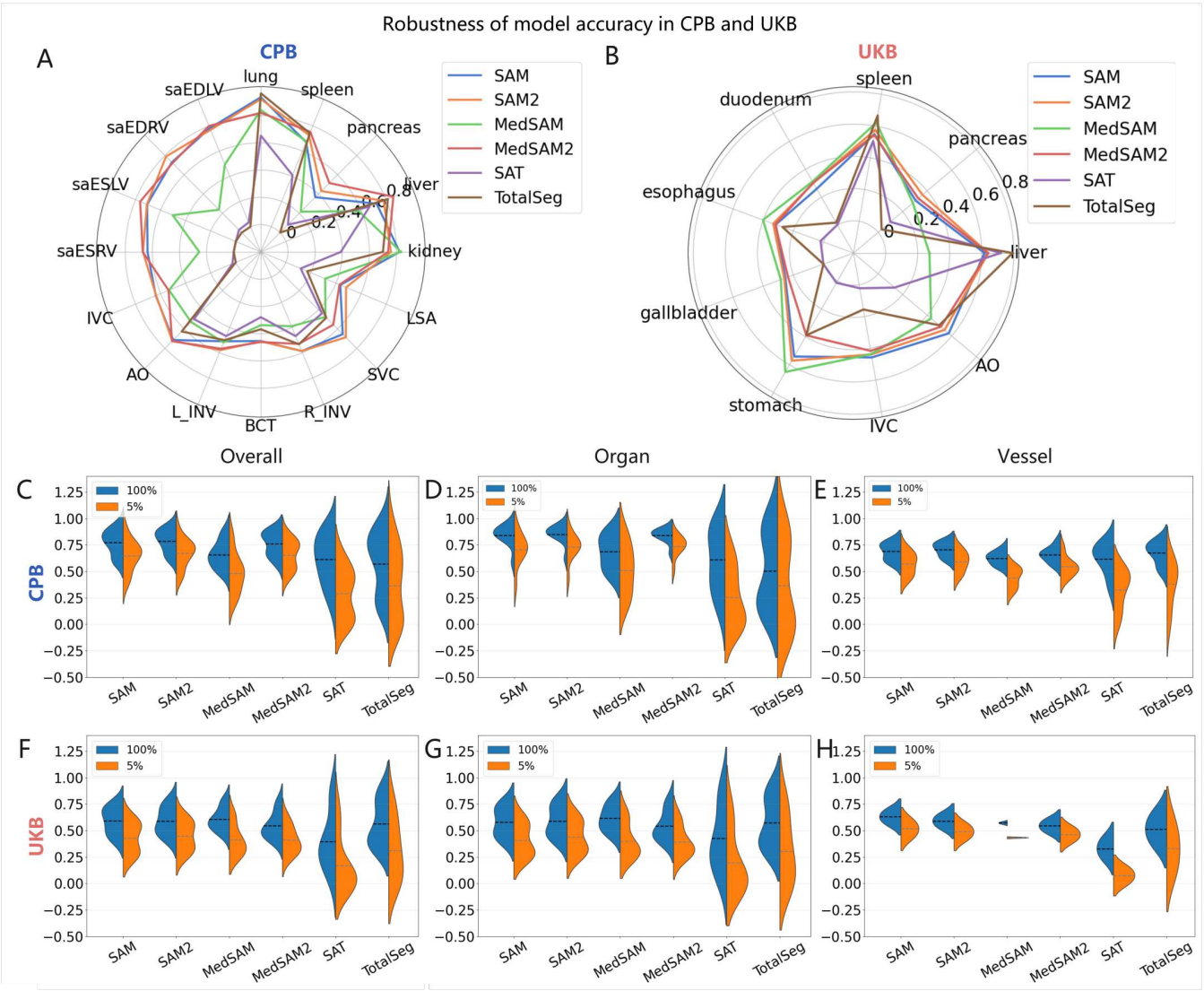


Fig. 4: **Robustness stress-testing uncovering the clinical brittleness of SFMs in challenging cases. A, B,** "Worst-case" performance profiles: Radar plots illustrating segmentation accuracy (DSC) on the most challenging 5% of cases for each anatomical structure in the **(A)** CPB and **(B)** UKB cohorts. While bbox-prompted models (e.g., SAM, SAM2) maintain stable performance, SAT and TotalSeg exhibit catastrophic collapse, particularly on complex cardiac and digestive targets. **C–H,** Performance distribution shifts: Violin plots comparing the DSC distributions between the most challenging 5% of cases (left) and the full cohort (right) for CPB **(C–E)** and UKB **(F–H)**. The substantial downward shift in median performance and the significant widening of interquartile ranges for SAT and TotalSeg underscore their inherent vulnerability to atypical anatomy and poor scan quality. These findings emphasize that average performance metrics mask critical unreliability in high-stakes clinical scenarios.

This task- and organ-dependent utility fragmentation extended seamlessly to the UKB cohort, which introduces additional complex digestive structures (e.g., duodenum, esophagus, gallbladder, stomach) absent from our in-house repository. In UKB multi-organ volumetry, while SAT sustained low relative volume errors (VER) on core abdominal structures like the liver, spleen, and pancreas, its accuracy collapsed precipitously on UKB-exclusive structures such as the stomach and esophagus. Concurrently, the SAM-family models exhibited volatile, highly inconsistent performance across these specialized organs. Ultimately, no individual baseline SFM was capable of maintaining a stable, low-VER profile across all investigated anatomical targets, reinforcing the critical necessity for a paradigm shift toward synergistic deployment models.

G. Synergistic model ensembles and knowledge distillation overcoming monolithic constraints

Given the localized, heavily compartmentalized strengths and structural vulnerabilities identified across standalone SFMs, we hypothesized that architectural synergy could systematically transcend the boundaries of individual monolithic

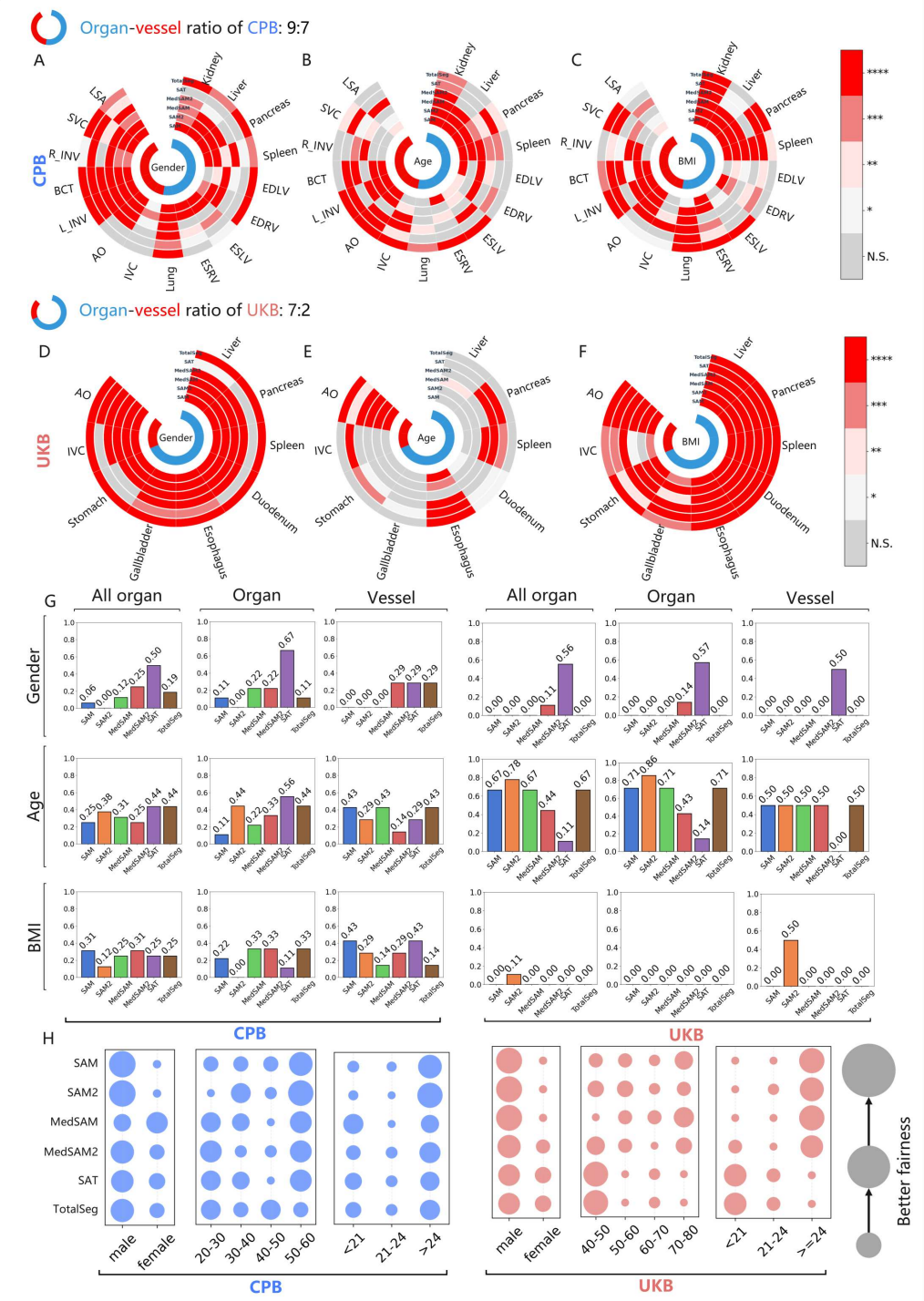


Fig. 5: **Comprehensive fairness auditing uncovering pervasive demographic disparities and systemic biases across SFMs. A–F,** Statistical significance of performance bias: Circular heatmaps illustrating subgroup disparities across 16 anatomical structures in the CPB **(A–C)** and UKB **(D–F)** cohorts. Statistical significance for each organ-model pair was determined by one-way ANOVA, with red gradients indicating a critical misalignment in performance across gender, age, and BMI. **G,** Aggregate fairness scores: The proportion of organs without significant performance bias (p ¿ 0.05 of ANOVA among groups) used to evaluate the overall fairness, the results highlight a pronounced accuracy–fairness trade-off in the CPB cohort where high-accuracy models (e.g., SAM2) exhibit the most frequent inequities. Notably, this fairness advantage is dataset-dependent, as evidenced by the diminished lead of the SAT model in the UKB age and BMI strata. **H,** Systematic bias portrait: Visualization of the direction of performance advantages, where bubble size corresponds to the number of organs for which a model achieved peak performance in a specific subgroup. The portrait reveals a consistent ”male-preference” and age-specific bias across most monolithic architectures, identifying potential risks for safe and equitable deployment in diverse clinical populations. Statistical significance: $*p \leq 0.05$; $**p \leq 0.01$; $***p \leq 0.001$; $****p \leq 0.0001$; N.S., not significant, lower significance means higher fairness.

networks. We validated this hypothesis by benchmarking five training-free post-processing fusion strategies alongside three multi-source knowledge distillation paradigms across both the in-house CPB and UKB cohorts. The empirical results demonstrate that the diagnostic dividends of model synergy are robustly generalizable and mathematically stable across divergent data distributions. While our primary analysis focuses on the DSC (Fig. 6) , complementary evaluations using the HD95 confirm highly congruent, cross-metric performance hierarchies (Supplementary Fig. S5).

Quantitative auditing demonstrates that both ensemble tracks consistently and decisively dominate every single standalone baseline SFM across all examined cohorts (Fig. 6). Within the training-free infrastructure, a simple voxel-wise average (AVG) fusion of the constituent models established a powerful baseline consensus. In the in-house dataset, AVG fusion achieved a superior mean DSC of 0.83 (compared to 0.79 for the top-performing standalone model) (Fig. 6A) , while simultaneously mitigating edge-case vulnerability and minimizing inter-organ performance volatility (Fig. 6C, E). Crucially, this collective intelligence track successfully dissolved the long-standing accuracy–fairness dilemma, optimizing demographic equity without compromising peak segmentational fidelity (Fig. 6G–I). This behavior replicated seamlessly within the independent UKB cohort, where AVG fusion elevated the mean DSC to 0.76 (surpassing the 0.70 achieved by MedSAM2) (Fig. 6B) , lowered inter-organ variance (Fig. 6D) , and preserved sub-population impartiality (Fig. 6J–L). These data confirm that principled fusion generalizes robustly beyond the training center boundaries.

Importantly, multi-source knowledge distillation yielded even more pronounced optimizations by synthesizing heterogeneous architectures into a compact student network. In the in-house dataset, the distilled student model (PUNet) set a new state-of-the-art benchmark with a mean DSC of 0.86, outperforming all heuristic training-free fusion operations (Fig. 6A). PUNet further secured exceptional generalization stability, compressing inter-organ standard deviation down to 0.07 compared to 0.11 for the best single standalone model, while substantially enhancing multi-attribute demographic fairness. This architectural dominance was mirrored in the external UKB cohort, where the distilled student frameworks (led by PUNet) consistently occupied peak positions in both aggregate accuracy (mean DSC 0.76) and variance reduction (Fig. 6B, D). Crucially, these student models successfully carried over and reinforced their fairness advantages within the UKB data distribution, establishing highly equitable sub-population profiles over individual baseline SFMs (Fig. 6J–L). Notably, the knowledge distillation process simultaneously achieved significant model size compression and reduced inference computational cost (Supplementary Table. S2), rendering the high-performance student models more feasible for deployment in resource-constrained clinical environments and reducing the overall cost of practical clinical application.

Beyond global macroscopic indices, we mapped the spatial error topographies of the distilled networks (Supplementary Fig. S6). These voxel-level visualizations reveal that knowledge distillation successfully minimizes the divergence of spatial error distributions, enforcing more consistent error magnitudes across varied anatomical regions and avoiding the wild boundary-to-interior fluctuations characteristic of baseline SFMs. Nevertheless, localized high-error zones remain identifiable in highly intricate anatomical sub-regions, indicating that while knowledge distillation successfully mitigates systematic geometric disparities, it does not completely immunize student networks against extreme structural edge cases.

This generalizable diagnostic superiority of collective intelligence is granularly validated by the downstream clinical metrics compiled in Table I. Within the cardiac domain, while individual baseline SFMs exhibited scattered and volatile compliance profiles—marking isolated victories (e.g., SAM2 in RVSV and MedSAM2 in RVEDV) but collapsing precipitously on other critical indices—the training-free AVG fusion systematically stabilized these erratic fluctuations, capturing the top performance tiers across nearly all functional parameters.

More profoundly, the multi-source knowledge-distilled PUNet student model achieved sweeping, near-total dominance across both clinical tasks and cohorts (Table I). PUNet fundamentally redefined the performance ceiling in cardiac volumetry, compressing error metrics to unprecedented margins (e.g., LVEDV to 4.14% and LVEF to 2.86%). This synergistic supremacy became absolute when expanding to abdominal and vascular structures. In the external UKB cohort, where monolithic models frequently defaulted to severe volumetric distortion on complex digestive targets, PUNet maintained unassailable diagnostic fidelity.

In summary, despite their heterogeneous design, both ensemble strategies produced a clinically more reliable and equitable performance profile than any individual SFM in both cohorts (Fig. 6M, N). This cross-dataset consistency highlights that the benefits of ensemble learning are not cohort-specific, but rather a general solution to the limitations of single models. The results suggest 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.

III. Discussion

This study establishes OrganSegBench, a rigorous multi-center benchmark anchored by two independent cohorts: the newly released CPB and the external UK Biobank. While raw UKB scans are accessible, the public repository lacks segmentation ground truths. To construct an unassailable evaluation yardstick, we invited radiologists to perform meticulous, slice-by-slice manual delineations across all 792 external subjects. Simultaneously, because CPB is introduced to the scientific community for the first time, it inherently eliminates the pervasive risk of training-data contamination. Thoroughly excavating this high-fidelity metadata (gender, age, BMI), we conducted a systematic, five-dimensional audit of six prominent SFMs across geometric accuracy, cross-task generalization, edge-case robustness, demographic fairness, and clinical utility.

Our multi-dimensional empirical analysis unmasked pervasive performance trade-offs, volatile model rankings, and

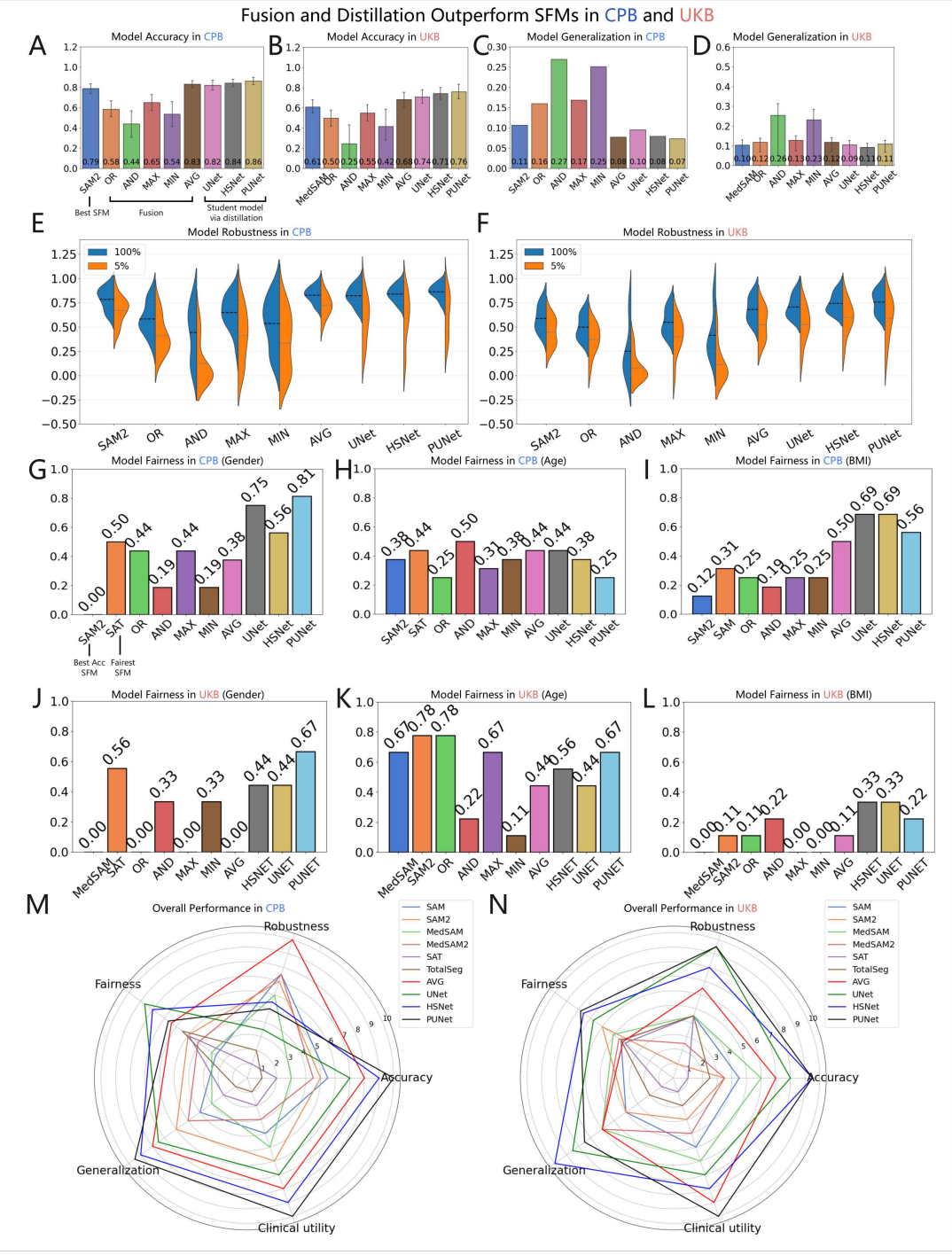


Fig. 6: **Synergistic ensemble learning and knowledge distillation architectures dominating individual monolithic SFMs. A, B,** Global segmentation accuracy: Overall DSC comparison on the (A) CPB and (B) UKB cohorts, demonstrating that the training-free average (AVG) fusion and multi-source knowledge-distilled student models (UNet, HSNet, PUNet) consistently deliver peak accuracy. **C, D,** Inter-organ performance variance: Standard deviation of DSC scores across all anatomical targets for **(C)** CPB and **(D)** UKB, showing that model synergy drastically minimizes performance volatility, signaling superior generalization stability. **E, F,** Stress-testing in worst-case scenarios: Violin plots showcasing DSC distributions between the full cohort (100%) and the most challenging 5% of edge cases for **(E)** CPB and **(F)** UKB. Synergistic methods, especially AVG fusion and PUNet, effectively safeguard boundary and region fidelity against catastrophic robustness collapse. **G–L,** Multi-attribute demographic fairness auditing: Aggregate fairness scores evaluating equity across gender **(G, J),** age **(H, K),** and BMI **(I, L)** for both datasets. The ensemble strategies systematically dissolve the long-standing accuracy–fairness dilemma by retaining high impartiality without sacrificing segmentation performance. **M, N,** Holistic performance synthesis: Radar summaries mapping comprehensive benchmarking parameters for **(M)** CPB and **(N)** UKB. The integrated profiles visually validate a major paradigm shift: collective intelligence via principled model synergy establishes a safer, more equitable, and superior clinical trajectory over traditional model scaling.

TABLE I: **Downstream clinical utility evaluation confronting monolithic SFMs and synergistic ensemble strategies across dual cohorts.** The table quantifies performance metrics across two distinct downstream clinical tasks: cardiac phenotype quantification (expressed as absolute error) and multi-organ volumetry (expressed as relative volume error [VER] %). Data are presented as mean (standard deviation). While individual monolithic foundation models present heavily compartmentalized and conflicting strengths: where bbox-prompted architectures excel in cardiac metrics but compromise general abdominal volumetry, the training-free AVG fusion and multi-source knowledge-distilled student models (UNet, HSNet, and PUNet) consistently deliver superior or highly competitive diagnostic fidelity across all anatomical targets in both CPB and UKB datasets. Bold values indicate the top three performing strategies within each phenotype or anatomical organ.

			Baseline-SFMs						**Ensemble strategies**			
Dataset	Category	Phenotype	SAM	SAM2	MedSAM	MedSAM2	SAT	TotalSeg	AVG fusion	UNet	HSNet	PUNet
CPB	**Cardiac function**	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)**
	Organ volume	kidney (%)	0.10(0.05)	0.14(0.05)	**0.05(0.03)**	0.30(0.03)	**0.05(0.04)**	0.24(0.05)	**0.06(0.03)**	0.06(0.04)	0.07(0.03)	0.08(0.02)
		liver (%)	0.18(0.10)	0.11(0.08)	0.11(0.07)	0.10(0.06)	**0.05(0.05)**	0.14(0.05)	0.06(0.05)	**0.04(0.06)**	**0.05(0.05)**	**0.02(0.03)**
		pancreas (%)	0.50(0.10)	0.51(0.09)	**0.26(0.16)**	0.56(0.09)	**0.30(0.20)**	0.53(0.21)	0.35(0.11)	**0.30(0.12)**	0.33(0.10)	**0.18(0.10)**
		spleen (%)	0.19(0.13)	0.13(0.09)	0.13(0.09)	0.23(0.08)	0.12(0.08)	0.21(0.06)	0.08(0.07)	**0.06(0.05)**	**0.07(0.06)**	**0.02(0.03)**
		lung (%)	0.06(0.02)	0.09(0.02)	0.12(0.03)	**0.05(0.03)**	**0.05(0.01)**	**0.04(0.00)**	**0.05(0.01)**	0.19(0.29)	0.20(0.28)	0.19(0.29)
	Vessel volume	IVC (%)	0.25(0.10)	0.33(0.08)	**0.11(0.08)**	0.49(0.08)	0.45(0.24)	0.39(0.19)	**0.13(0.07)**	0.22(0.12)	0.20(0.11)	**0.05(0.05)**
		AO (%)	0.32(0.04)	0.32(0.04)	0.24(0.06)	0.32(0.04)	0.18(0.09)	0.25(0.08)	**0.15(0.04)**	**0.16(0.08)**	0.23(0.05)	**0.06(0.04)**
		L_INV (%)	0.53(0.05)	0.45(0.05)	**0.11(0.08)**	0.49(0.04)	0.16(0.13)	**0.15(0.12)**	0.33(0.05)	0.28(0.11)	0.38(0.06)	**0.13(0.09)**
		BCT (%)	0.56(0.08)	0.54(0.07)	**0.14(0.10)**	0.59(0.06)	0.41(0.19)	**0.29(0.17)**	0.42(0.08)	0.38(0.15)	0.48(0.10)	**0.20(0.14)**
		R_INV (%)	0.36(0.08)	0.37(0.07)	**0.20(0.10)**	0.49(0.06)	0.36(0.14)	0.30(0.13)	**0.28(0.07)**	0.32(0.09)	0.34(0.07)	**0.06(0.05)**
		SVC (%)	0.35(0.07)	0.33(0.06)	0.28(0.08)	0.43(0.06)	0.27(0.12)	0.26(0.12)	**0.18(0.06)**	0.26(0.08)	**0.23(0.07)**	**0.07(0.04)**
		LSA (%)	0.44(0.12)	0.50(0.07)	**0.14(0.10)**	0.60(0.06)	0.56(0.14)	0.59(0.13)	**0.35(0.08)**	0.43(0.13)	0.47(0.09)	**0.23(0.14)**
UKB	**Organ volume**	liver (%)	0.22(0.07)	0.17(0.07)	0.22(0.07)	0.08(0.06)	**0.05(0.05)**	0.12(0.03)	0.08(0.05)	**0.05(0.03)**	0.08(0.05)	**0.01(0.01)**
		pancreas (%)	0.36(0.08)	0.49(0.06)	0.36(0.08)	0.38(0.08)	**0.20(0.15)**	0.62(0.17)	**0.19(0.07)**	0.32(0.08)	0.31(0.06)	**0.14(0.08)**
		spleen (%)	0.26(0.08)	0.23(0.07)	0.26(0.08)	0.22(0.05)	**0.07(0.08)**	0.14(0.05)	**0.12(0.05)**	0.14(0.05)	0.14(0.05)	**0.02(0.04)**
		duodenum (%)	0.37(0.13)	0.49(0.10)	0.37(0.13)	0.47(0.07)	0.54(0.20)	0.63(0.21)	**0.22(0.10)**	**0.29(0.12)**	0.33(0.11)	**0.14(0.10)**
		esophagus (%)	0.39(0.06)	0.54(0.05)	0.39(0.06)	0.58(0.04)	0.64(0.16)	0.55(0.06)	**0.26(0.06)**	**0.36(0.05)**	0.38(0.05)	**0.12(0.04)**
		gallbladder (%)	**0.30(0.16)**	0.42(0.14)	**0.30(0.16)**	0.49(0.12)	0.47(0.27)	0.57(0.31)	**0.20(0.14)**	**0.30(0.18)**	**0.30(0.18)**	**0.22(0.19)**
		stomach (%)	0.14(0.08)	0.23(0.07)	0.14(0.08)	0.13(0.08)	0.60(0.20)	0.20(0.11)	**0.04(0.03)**	0.05(0.05)	**0.04(0.04)**	**0.03(0.04)**
	Vessel volume	IVC (%)	0.26(0.06)	0.40(0.05)	0.26(0.06)	0.47(0.05)	0.67(0.14)	0.57(0.12)	**0.14(0.05)**	**0.22(0.06)**	0.23(0.05)	**0.06(0.03)**
		AO (%)	0.25(0.04)	0.33(0.04)	0.25(0.04)	0.37(0.04)	0.55(0.13)	0.38(0.04)	**0.12(0.04)**	**0.15(0.04)**	0.17(0.04)	**0.04(0.03)**
		Overall (%)	0.28(0.08)	0.37(0.13)	0.28(0.08)	0.35(0.17)	0.42(0.25)	0.42(0.21)	**0.15(0.07)**	**0.21(0.12)**	0.22(0.12)	**0.09(0.07)**

cross-cohort performance reversals that challenge prevailing dogmas in medical AI, revealing a critical disconnect between SFM's theoretical promise and their practical readiness for clinical deployment [30]. First, the surprising superiority of the generalist SAM over its domain-specialized counterpart, MedSAM, within the CPB dataset suggests that current localized medical fine-tuning risks introducing catastrophic representation erosion, potentially fracturing the highly robust latent spaces learned from diverse, large-scale pre-training distributions [19], [31]. Second, by moving beyond abstract global indices, our 3D organ point-cloud visualization exposed distinct, paradigm-specific error topographies entirely invisible to standard metrics like DSC. Third, by simulating downstream diagnostic pipelines, cardiac function analysis and multi-organ volumetry [28], [29], we demonstrate that foundation model utility is strictly task-dependent, underscoring the critical need for context-aware model selection over isolated bench metrics.

The most significant conceptual advance of this work is the definitive demonstration that ensemble learning strategies can systematically overcome the hard anatomical and ethical boundaries constraining individual monolithic networks. The uncompromised superiority demonstrated by both the training-free average (AVG) fusion and multi-source knowledge distillation frameworks across almost all evaluation vectors represents a profound paradigm shift. It directly challenges the dominant, computationally extravagant trajectory of AI development, which disproportionately prioritizes the brute-force scaling of ever-larger monolithic models [32]–[34].

Crucially, our framework offers a flexible deployment matrix: the voxel-averaging ensemble requires zero training overhead and is uniquely optimized for low-resource environments or highly heterogeneous datasets. Conversely, multi-source knowledge distillation successfully synthesizes the complementary latent knowledge of diverse experts into a compact, highly efficient student network (PUNet), making it ideal for massive data centers where post-training throughput, localized compression, and task-specific optimization are paramount. Together, these approaches highlight the vast, untapped potential of collective intelligence, charting a more pragmatic path forward: model synergy over model scale [35].

However, this study has several limitations that need to be acknowledged. First, our demographic fairness auditing was bound by the available cohort metadata to gender, age, and BMI; other vital socio-demographic, geographic, and scanner-hardware attributes should be incorporated in future iterations. Second, the evaluation of bbox-prompted models relied on manual bounding boxes curated by radiologists; a fully automated, prompt-free closed-loop system was not deployed within this specific work [36]. Third, although PUNet successfully minimized the divergence of spatial error distributions, localized high-error zones persist within highly intricate anatomical corners or fine-grained vascular junctions.

In conclusion, this study provides a vital reality check on medical segmentation foundation models. Even state-of-the-art SFMs exhibit severe weaknesses in robustness under stress and demographic equity, vulnerabilities completely obscured

by conventional accuracy metrics alone. Crucially, our results illuminate a mathematically rigorous way forward: ensemble learning strategies systematically and comprehensively outperform monolithic systems across every examined dimension. This principle of synergy over scale marks a major paradigm shift for medical AI, transitioning the community away from erratic standalone systems toward the intelligent synthesis of complementary model ecosystems. OrganSegBench provides a clear, actionable roadmap for engineering the next generation of medical AI: effective, safe, equitable, and ultimately worthy of clinical trust.

IV. Methods

A. The OrganSegBench Dataset

1) China PhenoBank: To evaluate the performance of SFMs and eliminate potential data leakage from pre-training corpora, we utilized the data resource from PhenoBank [37], [38] (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 and 3 scanners. 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 mm^2, 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 mm^2, spatial resolution of 0.7 × 0.7 mm^2, and TR/TE of 3.42/1.50 ms.

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/m^2). Further details on demographic metadata per subject and inclusion/exclusion criteria are provided in the supplementary Table. S1 and Supplementary Fig. S1A.

2) UKBiobank: To construct a standardized external benchmark for systematic model evaluation and validate the cross-cohort robustness of vanilla segmentation models and knowledge distillation frameworks under heterogeneous populations and multi-center imaging protocols, we adopted the population-based UKB cohort (https://www.ukbiobank.ac.uk/). The UKB imaging substudy acquired multimodal imaging data from 100,000 participants. For this study, we specifically selected a representative subset of 792 participants with complete whole-body MRI data acquired using the whole Dixon sequence, aiming to ensure the balance of population attribute distribution. The selected subset included demographic information for fairness evaluation: gender (365 males, 427 females), age (ranging from 40 to 80 years, consistent with UKB recruitment criteria), and BMI (ranging from 16.3 to 42.1 kg/m^2). Further details on UKB imaging protocols and data processing pipelines are provided in the UKB imaging documentation.

3) Data Annotation: For each of the 701 subjects in CPB, a total of 16 anatomical structures (9 large organs and 7 major vessels) were manually segmented to create ground-truth segmentation 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. S1B.

For the 792 included subjects in UKB, segmentation annotations of 9 anatomical organs were obtained through a semi-automatic workflow, aiming to reduce annotation costs while ensuring annotation accuracy. Specifically, we first performed pre-segmentation on all images using VIBESegmentator [39], a reliable automatic segmentation tool optimized for MRI modalities. Subsequently, the pre-segmented masks were subjected to manual quality control and refinement by the same five radiologists, to correct segmentation errors (e.g., boundary inaccuracies and missing regions) and ensure the reliability of annotations.

B. Organ Segmentation using SFMs

We benchmarked six prominent SFMs, 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 ($\mathcal{M}_{\text{SAM}}$), MedSAM ($\mathcal{M}_{\text{MedSAM}}$), SAM2 ($\mathcal{M}_{\text{SAM2}}$), MedSAM2 ($\mathcal{M}_{\text{MedSAM2}}$), SAT ($\mathcal{M}_{\text{SAT}}$), and TotalSegmentator ($\mathcal{M}_{\text{TotalSeg}}$). All models were used with their original pre-trained weights.

- **Bbox-Prompted Models:** $\mathcal{M}_{\text{SAM}}$, $\mathcal{M}_{\text{MedSAM}}$, $\mathcal{M}_{\text{SAM2}}$, and $\mathcal{M}_{\text{MedSAM2}}$ 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 $\hat{Y}^N_{i,j,o} = \mathcal{M}_N(x_{i,j}, B_{i,j,o})$, where $N \in \{1, 2, 3, 4\}$ corresponds to the four models.
- **Text-Prompted Model:** $\mathcal{M}_{\text{SAT}}$ 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 $\hat{Y}^{\text{SAT}}_{i,\text{mod},o} = \mathcal{M}_{\text{SAT}}(X_{i,\text{mod}}, \text{"mod"}, \text{"o"})$. We used the MONAI library to restore output dimensions to match the original input.

- **Prompt-Free Model:** $\mathcal{M}_{\text{TotalSeg}}$ is a 3D model that requires no external prompts. It was applied directly using its default configuration, with the output defined as $\hat{Y}^{\text{TotalSeg}}_{i,\text{mod},o} = \mathcal{M}_{\text{TotalSeg}}(X_{i,\text{mod}})$.

C. Multi-Dimensional Performance Evaluation

1) Accuracy and Generalization: We evaluated volumetric and boundary accuracy using the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95), respectively. The average performance for a given model m on an organ o across all samples was recorded as $DSC^{\text{all}}_{o,m}$ and $HD95^{\text{all}}_{o,m}$. 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 IV-D). As for generalization, we quantified the variance of accuracy among organs via standard deviation.

2) 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, $S^{\text{DSC, 5\%}}_{o,m}$, and the highest 5% HD95 values, $S^{\text{HD95, 5\%}}_{o,m}$. The average performance on these subsets, $\overline{DSC}^{5\%}_{o,m}$ and $\overline{HD95}^{5\%}_{o,m}$, was calculated to quantify robustness. Performance degradation was measured as the difference between the overall average and this worst-case average.

3) 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.

4) Clinical Utility: We evaluated the applicability of the models for downstream clinical tasks in in house dataset and UKB. For cardiac analysis in in house dataset, we computed nine phenotypes, including LVEDV (Left Ventricle End-Diastolic Volume), RVEDV (Right Ventricle End-Diastolic Volume), LVESV (Left Ventricle End-Systolic Volume), RVESV (Right Ventricle End-Systolic Volume), LVSV (Left Ventricular Stroke Volume), RVSV (Right Ventricular Stroke Volume), LVEF (Left Ventricular Ejection Fractions), RVEF (Right Ventricular Ejection Fractions) and LVCO (Left Ventricular Cardiac Output), from the model-generated masks and calculated the **absolute error** against phenotypes derived from ground-truth masks. The details of the cardiac phenotypes computing can be found in supplementary algorithm 1.

For non-cardiac organs in the CPB cohort and all investigated organs in the UKB cohort, only organ volume was adopted as the evaluation phenotype. For these 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**, $VER = \frac{|V_{\text{GT}} - \hat{V}|}{V_{\text{GT}}}$. More sophisticated functional and clinical phenotypes cannot be directly extracted merely from static anatomical segmentation masks, as their quantification requires additional dynamic imaging sequences, hemodynamic information, or specialized clinical measurements that are not available in the current CPB and UKB datasets.

D. 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 algorithm 2 (see *https://github.com/lq2233/OrganSegBench*).

E. Ensemble Learning Strategies

1) 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 $\hat{Y}^{N}_{i,j,o}(p) \in \{0, 1\}$. The fused mask voxel is $\hat{Y}^{\text{ensemble}}_{i,j,o}(p)$.

- **OR / AND:** Logical operations on the six masks. The output voxel is $\hat{Y}^{\text{OR}}_{i,j,o}(p) = \cup_N(\hat{Y}^{N}_{i,j,o}(p))$ for OR, and $\hat{Y}^{\text{AND}}_{i,j,o}(p) = \cap_N(\hat{Y}^{N}_{i,j,o}(p))$ for AND.
- **MIN / MAX:** Voxel-count-based selection. Let $C^{N}_{i,j,o} = \sum_p \hat{Y}^{N}_{i,j,o}(p)$ 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 $\hat{Y}^{\text{MIN}}_{i,j,o} = \hat{Y}^{N^*}_{i,j,o}$ 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).

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

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

2) Multi-Source Knowledge Distillation: While Segment Anything Models (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 [40] and HSNet [41], where UNet is CNN-based model and HSNet is a Transformer-driven model (using PvT [42]).

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 $\hat{Y}_{i,j,o}^{\text{AVG}}$ given the input 2D slice image X_i. The loss function is a combination of Intersection over Union (IoU) loss [43] and Binary Cross-Entropy (BCE) loss [44] (in Eq. 2).

$$\mathcal{L}(X_i, \hat{Y}_{i,j,o}^{\text{AVG}}) = \sum_{m=1}^{M} [\mathcal{L}_{IoU}(S_m(X_i), \hat{Y}_{i,j,o}^{\text{AVG}}) + \mathcal{L}_{BCE}(S_m(X_i), \hat{Y}_{i,j,o}^{\text{AVG}})], \tag{2}$$

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: $S_{inference}(X_i) = \sigma(S(X_i))$. 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 $\mathbb{R}^N$ ($N = 6$ in our implementation, corresponding to the number of SFMs) to represent segmentation variability. The full model consists of a prior network $P(\mathbf{z}|X_i)$, a posterior network $Q(\mathbf{z}|X_i, Y_i)$, and a U-Net backbone. *(a) Prior Network:* Generates a Gaussian distribution over the latent space, conditioned on the input image X_i: $\mathbf{z} \sim P(\cdot|X_i) = \mathcal{N}(\boldsymbol{\mu}_{prior}(X_i;\omega), \boldsymbol{\Sigma}_{prior}(X_i;\omega))$, 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 Y_i: $\mathbf{z} \sim Q(\cdot|X_i, Y_i) = \mathcal{N}(\boldsymbol{\mu}_{post}(X_i, Y_i;\nu), \boldsymbol{\Sigma}_{post}(X_i, Y_i;\nu))$, where ν denotes the parameters of the posterior network. *(c) Segmentation Generation:* A latent vector $\mathbf{z}$, 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: $S(X_i, \mathbf{z}) = f_{comb}(f_{U-Net}(X_i;\theta), \mathbf{z};\psi)$ 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: $Y_i = \sum_{k=1}^{6} \mathbb{I}(K = k) Y_{i,F_k}$, where $\mathbb{I}(\cdot)$ denotes the indicator function.

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

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

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 $P(\mathbf{z}|X_i)$ to sample N(N=10 in our study) latent vectors, $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N\}$, where $\mathbf{z}_n \sim P(\cdot|X_i)$ for $n = 1, \ldots, N$. Each latent vector $\mathbf{z}_n$ is then used to generate a corresponding segmentation mask $S(X_i, \mathbf{z}_n)$. The final segmentation mask, $S_{final}(X_i)$, is obtained by averaging these individual masks: $S_{final}(X_i) = \frac{1}{N}\sum_{n=1}^{N} S(X_i, \mathbf{z}_n)$.

All models are trained with unified training epochs and batch size, and the detailed training hyperparameters are presented in Supplementary Table S3.

F. Data availability

The China PhenoBank (CPB) dataset used in this study is available via PhenoBank (*https://www.phenobank.org.cn/*). Full access to the data can be obtained upon submitting a formal request and securing approval through the PhenoBank portal. The external validation data were obtained from the UK Biobank Resource (https:// www.ukbiobank.ac.uk/) under Application Number 96511. Qualified researchers can gain access to the data by submitting a formal application via its official access management system.

G. Code availability

The source code for training and inference of multi-source knowledge distillation models, as well as computing and visualizing spatial segmentation error distributions has been made publicly available at https://github.com/lq2233/OrganSegBench.

V. Acknowledgment

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). The computations in this research were performed using the CFFF platform of Fudan University.

REFERENCES

[1] Y. Sun, L. Wang, G. Li, W. Lin, and L. Wang, "A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks," *Nature Biomedical Engineering*, vol. 9, no. 4, pp. 521–538, 2025.

[2] J. Wang, K. Wang, Y. Yu, Y. Lu, W. Xiao, Z. Sun, F. Liu, Z. Zou, Y. Gao, L. Yang *et al.*, "Self-improving generative foundation model for synthetic medical image generation and clinical applications," *Nature Medicine*, vol. 31, no. 2, pp. 609–617, 2025.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[4] S. Yan, Z. Yu, C. Primiero, C. Vico-Alonso, Z. Wang, L. Yang, P. Tschandl, M. Hu, L. Ju, G. Tan *et al.*, "A multimodal vision foundation model for clinical dermatology," *Nature Medicine*, pp. 1–12, 2025.

[5] J. Ma and B. Wang, "Towards foundation models of biological image segmentation," *Nature Methods*, vol. 20, no. 7, pp. 953–955, 2023.

[6] J. Ma, S. Kim, F. Li, M. Baharoon, R. Asakereh, H. Lyu, and B. Wang, "Segment anything in medical images and videos: Benchmark and deployment," *arXiv preprint arXiv:2408.03322*, 2024.

[7] Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Chen, C. Chen *et al.*, "Segment anything model for medical images?" *Medical Image Analysis*, vol. 92, p. 103061, 2024.

[8] S. Pai, D. Bontempi, I. Hadzic, V. Prudente, M. Sokač, T. L. Chaunzwa, S. Bernatz, A. Hosny, R. H. Mak, N. J. Birkbak *et al.*, "Foundation model for cancer imaging biomarkers," *Nature machine intelligence*, vol. 6, no. 3, pp. 354–367, 2024.

[9] A. Archit, L. Freckmann, S. Nair, N. Khalid, P. Hilt, V. Rajashekar, M. Freitag, C. Teuber, M. Spitzner, C. Tapia Contreras *et al.*, "Segment anything for microscopy," *Nature Methods*, vol. 22, no. 3, pp. 579–591, 2025.

[10] Z. Huai, H. Tang, Y. Li, Z. Chen, and X. Li, "Leveraging segment anything model for source-free domain adaptation via dual feature guided auto-prompting," *arXiv preprint arXiv:2505.08527*, 2025.

[11] C. Zhang, J. Cho, F. D. Puspitasari, S. Zheng, C. Li, Y. Qiao, T. Kang, X. Shan, C. Zhang, C. Qin *et al.*, "A survey on segment anything model (sam): Vision foundation model meets prompt engineering," *arXiv preprint arXiv:2306.06211*, 2023.

[12] S. He, R. Bao, J. Li, J. Stout, A. Bjornerud, P. E. Grant, and Y. Ou, "Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets," *arXiv preprint arXiv:2304.09324*, 2023.

[13] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: an experimental study," *Medical Image Analysis*, vol. 89, p. 102918, 2023.

[14] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[15] Z. Zhao, Y. Zhang, C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "One model to rule them all: Towards universal segmentation for medical images with text prompts," *arXiv preprint arXiv:2312.17183*, 2023.

[16] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang *et al.*, "Totalsegmentator: robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, p. e230024, 2023.

[17] M. A. Ricci Lara, R. Echeveste, and E. Ferrante, "Addressing fairness in artificial intelligence for medical imaging," *nature communications*, vol. 13, no. 1, p. 4581, 2022.

[18] R. J. Chen, J. J. Wang, D. F. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai, and F. Mahmood, "Algorithmic fairness in artificial intelligence for medicine and healthcare," *Nature biomedical engineering*, vol. 7, no. 6, pp. 719–742, 2023.

[19] Q. Li, Y. Zhang, Y. Li, J. Lyu, M. Liu, L. Sun, M. Sun, Q. Li, W. Mao, X. Wu *et al.*, "An empirical study on the fairness of foundation models for multi-organ image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 432–442.

[20] J. Ma, Y. Xu, F. Zhou, Y. Wang, C. Jin, Z. Guo, J. Wu, O. K. Tang, H. Zhou, X. Wang *et al.*, "Pathbench: A comprehensive comparison benchmark for pathology foundation models towards precision oncology," *arXiv preprint arXiv:2505.20202*, 2025.

[21] M. Huang, H. Han, H. Wang, L. Li, Y. Zhang, and U. A. Bhatti, "A clinical decision support framework for heterogeneous data sources," *IEEE journal of biomedical and health informatics*, vol. 22, no. 6, pp. 1824–1833, 2018.

[22] G. Kumar, S. Basri, A. A. Imam, S. A. Khowaja, L. F. Capretz, and A. O. Balogun, "Data harmonization for heterogeneous datasets: a systematic literature review," *Applied Sciences*, vol. 11, no. 17, p. 8275, 2021.

[23] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek *et al.*, "Metrics reloaded: recommendations for image analysis validation," *Nature methods*, vol. 21, no. 2, pp. 195–212, 2024.

[24] Q. Li, Y. Zhang, S. Yang, Q. Li, Z. Wang, J. Liu, H. Zhang, S. Wang, and C. Wang, "From generalist to specialist: Distilling a mixture of foundation models for domain-specific medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 197–206.

[25] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[26] J. Zhu, A. Hamdi, Y. Qi, Y. Jin, and J. Wu, "Medical sam 2: Segment medical images as video via segment anything model 2," *arXiv preprint arXiv:2408.00874*, 2024.

[27] A. Balendran, C. Beji, F. Bouvier, O. Khalifa, T. Evgeniou, P. Ravaud, and R. Porcher, "A scoping review of robustness concepts for machine learning in healthcare," *npj Digital Medicine*, vol. 8, no. 1, p. 38, 2025.

[28] A. L. Colunga, K. G. Kim, N. P. Woodall, T. F. Dardas, J. H. Gennari, M. S. Olufsen, and B. E. Carlson, "Deep phenotyping of cardiac function in heart transplant patients using cardiovascular system models," *The Journal of physiology*, vol. 598, no. 15, pp. 3203–3222, 2020.

[29] M. U. Bevilacqua, C. J. Hague, A. Romann, H. Sheitt, D. M. Vasilescu, T. W. Yi, and A. Levin, "Ct of kidney volume in autosomal dominant polycystic kidney disease: accuracy, reproducibility, and radiation dose," *Radiology*, vol. 291, no. 3, pp. 660–667, 2019.

[30] T. Grote, T. Freiesleben, and P. Berens, "Foundation models in healthcare require rethinking reliability," *Nature Machine Intelligence*, vol. 6, no. 12, pp. 1421–1423, 2024.

[31] H. Zheng, L. Shen, A. Tang, Y. Luo, H. Hu, B. Du, Y. Wen, and D. Tao, "Learning from models beyond fine-tuning," *Nature Machine Intelligence*, vol. 7, no. 1, pp. 6–17, 2025.

[32] K. Zhang, R. Zhou, E. Adhikarla, Z. Yan, Y. Liu, J. Yu, Z. Liu, X. Chen, B. D. Davison, H. Ren *et al.*, "A generalist vision–language foundation model for diverse biomedical tasks," *Nature Medicine*, vol. 30, no. 11, pp. 3129–3141, 2024.

[33] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[34] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena *et al.*, "Towards generalist biomedical ai," *Nejm Ai*, vol. 1, no. 3, p. AIoa2300138, 2024.

[35] G. Wang, J. Ye, J. Cheng, T. Li, Z. Chen, J. Cai, J. He, and B. Zhuang, "Sam-med3d-moe: Towards a non-forgetting segment anything model via mixture of experts for 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 552–561.

[36] Y. Li, L. Zhang, Y. Liang, and P. Xie, "Am-sam: Automated prompting and mask calibration for segment anything model," *arXiv preprint arXiv:2410.09714*, 2024.

[37] C. Wang, Z. Shi, Y. Li, X. Xia, X. Kuang, S. Qian, L. Xue, L. Lan, Y. Wu, N. Zhang *et al.*, "Protocol for brain magnetic resonance imaging and extraction of imaging-derived phenotypes from the china phenobank project," *Phenomics*, vol. 3, no. 6, pp. 642–656, 2023.

[38] C. Wang, S. Wang, S. Hua, R. Li, Y. Li, Z. Shi, K. Feng, L. Lan, M. Liu, X. Kuang *et al.*, "A protocol for body mri/ct and extraction of

imaging-derived phenotypes (idps) from the china phenobank project," *Phenomics*, vol. 4, no. 6, pp. 594–616, 2024.

[39] R. Graf, P. Platzek, E. O. Riedel, C. Ramschütz, S. Starck, H. K. Möller, M. Atad, H. Völzke, R. Bülow, C. O. Schmidt *et al.*, "Vibesegmentator: full body mri segmentation for the nako and uk biobank," *European Radiology*, pp. 1–15, 2025.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[41] W. Zhang, C. Fu, Y. Zheng, F. Zhang, Y. Zhao, and C.-W. Sham, "Hsnet: A hybrid semantic network for polyp segmentation," *Computers in biology and medicine*, vol. 150, p. 106173, 2022.

[42] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[43] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.

[44] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

Supplementary Files

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

- Supplementary.pdf