Supplementary Materials

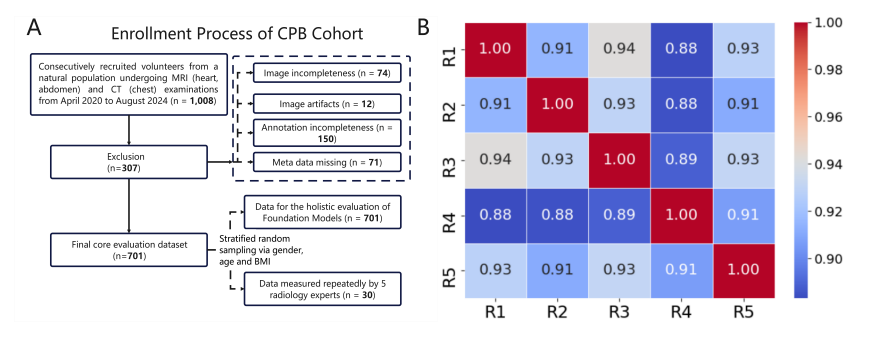

Figure S1: **Characterization and multi-stage quality control of the ChinaPhenoBank (CPB) cohort.A.**Flowchart of subject enrollment and sub-cohort stratification. **B.** Inter-rater segmentation reliability matrix: Correlation heatmap quantifying the overall annotation consistency among five independent radiologists (R1–R5) evaluated on the reliability sub-cohort (N=30). The pairwise values represent the Kappa coefficients calculated and averaged across all 16 target organs and vessels, demonstrating exceptional and highly reproducible diagnostic consensus (overall mean Kappa = 0.94).

Table S1: **Dataset (CPB) attribute description with population characteristics.**

Attribute	Total	Male	Female
Samples (n)	701	289	412
Age (years; mean, SD)	35.09 (11.76)	35.09 (11.42)	35.09 (11.99)
Height (m; mean, SD)	1.66 (0.09)	1.73 (0.06)	1.61 (0.06)
Weight (kg; mean, SD)	63.65 (12.06)	71.57 (10.52)	58.24 (9.83)
BMI (kg/m^2; mean, SD)	23.07 (3.47)	23.79 (3.13)	22.56 (3.61)
Kidney volume (mL, mean, SD)	282.92 (50.00)	310.05 (47.44)	263.97 (42.41)
Liver volume (mL, mean, SD)	1243.27 (279.96)	1402.82 (262.00)	1131.25 (234.23)
Pancreas volume (mL, mean, SD)	62.91 (16.97)	70.11 (17.04)	57.83 (14.97)
Spleen volume (mL, mean, SD)	195.26 (73.32)	230.73 (78.73)	170.39 (57.56)
EDLV volume (mL, mean, SD)	135.01 (32.94)	153.76 (35.91)	121.91 (22.85)
ESLV volume (mL, mean, SD)	53.21 (18.73)	60.77 (21.18)	47.92 (14.64)
EDRV volume (mL, mean, SD)	142.23 (40.30)	165.21 (43.99)	126.16 (27.87)
ESRV volume (mL, mean, SD)	71.43 (26.11)	85.29 (27.71)	61.68 (19.82)
IVC (mL, mean, SD)	39.11 (17.34)	42.92 (18.65)	36.38 (15.81)
Lung volume (mL, mean, SD)	4283.11 (1046.36)	5084.58 (915.66)	3723.23 (714.95)
AO (mL, mean, SD)	114.82 (37.02)	133.30 (36.74)	101.98 (31.26)
L_INV (mL, mean, SD)	8.93 (2.92)	10.47 (3.04)	7.86 (2.30)
BCT (mL, mean, SD)	2.66 (1.44)	3.25 (1.17)	2.25 (1.47)
R_INV (mL, mean, SD)	4.19 (1.40)	4.79 (1.55)	3.77 (1.12)
SVC (mL, mean, SD)	11.22 (3.23)	12.90 (3.52)	10.05 (2.40)
LSA (mL, mean, SD)	2.06 (0.79)	2.52 (0.80)	1.73 (0.60)

SD: standard deviation.
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;
BCT, Brachiocephalic Trunk; R_INV, right brachiocephalic vein;
SVC, Superior Vena Cava; LSA, Left Subclavian Artery.

Table S2: Model Size and Inference Cost for 512x512 Input.

Models	SAT	SAM	MedSAM	SAM2	MedSAM2	TotalSeg	UNet	HSNet	PUNet
Param(M)	878.59	635.63	90.48	216.90	31.42	49.08	2.26	29.85	5.00
FLOPs(G)	1090.18	2736.62	371.99	2813.98	106.72	70.42	55.65	23.15	96.95

Table S3: Training hyperparameters of different student models

Model	Epochs	Batch Size	Learning Rate	Weight Decay	Optimizer
Probabilistic U-Net	1000	20	1×10^{-5}	1×10^{-6}	AdamW
U-Net	1000	20	1×10^{-4}	1×10^{-5}	AdamW
HSNet	1000	20	1×10^{-4}	1×10^{-5}	AdamW

Algorithm 1: Cardiac Phenotype Computation

Input: V_{EDLV}, V_{EDRV}, V_{ESLV}, V_{ESRV} as the volume (mL) of the region EDLV, EDRV, ESLV and ESRV of cardiac respectively.

Output: 9 phenotypes, L/RV of Left/Right Ventricle

LVEDV = V_{EDLV}, LVESV = V_{ESLV};
RVEDV = V_{EDRV}, RVESV = V_{ESRV};
LVSV = LVEDV - LVESV;
RVSV = RVEDV - RVESV;
LVEF = LVSV/LVEDV $\times$ 100;
RVEF = RVSV/RVEDV $\times$ 100;
LVCO = LVSV $\times$ alpha $\times$ 10^{-3};
alpha = 60.0/duration/cycle;

Where:

LVEDV: Left Ventricular End-Diastolic Volume
LVESV: Left Ventricular End-Systolic Volume
RVEDV: Right Ventricular End-Diastolic Volume
RVESV: Right Ventricular End-Systolic Volume
LVSV/RVSV: Stroke Volume of Left/Right Ventricle
LVEF/RVEF: Ejection Fraction of Left/Right Ventricle
LVCO: Left Ventricular Cardiac Output
duration: Cardiac cycle duration (s)
cycle: Number of cardiac cycles

Algorithm 2: Spatial segmentation error distribution visualization among sample groups

Input: Organ $o \in$ {kidney, liver, spleen, lung, AO};
Dataset $D = \{d_j\}, j = 1, \ldots, 701$;
Segmentation results S_j of Segmentation Foundation Models with annotations L_j

Output: Overall distribution of segmentation error of Segmentation Foundation Models for organ o in D

Initialization: Select one sample's mask from dataset as reference Ref_o;
foreach *sample* $d_j \in D$ **do**
  Transfer S_j, L_j into STL format, then into point cloud data SP_j, LP_j;
  Compute point-to-point distance $Distance_{SL}$ from SP_j to LP_j;
  Compute point-to-point distance $Distance_{LS}$ from LP_j to SP_j;
  $P_{95} \leftarrow \text{Percentile}([Distance_{SL}, Distance_{LS}], 95)$;
  $Distance_{SL}[Distance_{SL} > P_{95}] \leftarrow P_{95}$;
  $Distance_{LS}[Distance_{LS} > P_{95}] \leftarrow P_{95}$;
  Initialize $Count_{point} \leftarrow 1$;
  foreach *point* $point_{SP} \in SP_j$ **do**
    $point_{SPtoLP} \leftarrow \text{find\_nearest}(LP_j, point_{SP})$;
    $Distance_{SL}[point_{SPtoLP}] \leftarrow Distance_{SL}[point_{SPtoLP}] + Distance_{LS}[point_{SP}]$;
    $Count[point_{SPtoLP}] \leftarrow Count[point_{SPtoLP}] + 1$;
  end
  foreach *point* $point_{LP} \in LP_j$ **do**
    $Distance_{SL}[point_{LP}] \leftarrow Distance_{SL}[point_{LP}]/Count[point_{LP}]$;
    $IDX_{LPtoRef} \leftarrow \text{find\_nearest}(Ref_o, point_{LP})$;
    $Ref_o[IDX_{LPtoRef}].distance \leftarrow Distance_{SL}[point_{LP}]$;
  end
end
foreach *point* $point_{Ref} \in Ref_o$ **do**
  $Ref_o[IDX_{LPtoRef}].distance \leftarrow Ref_o[IDX_{LPtoRef}].distance/701$;
end

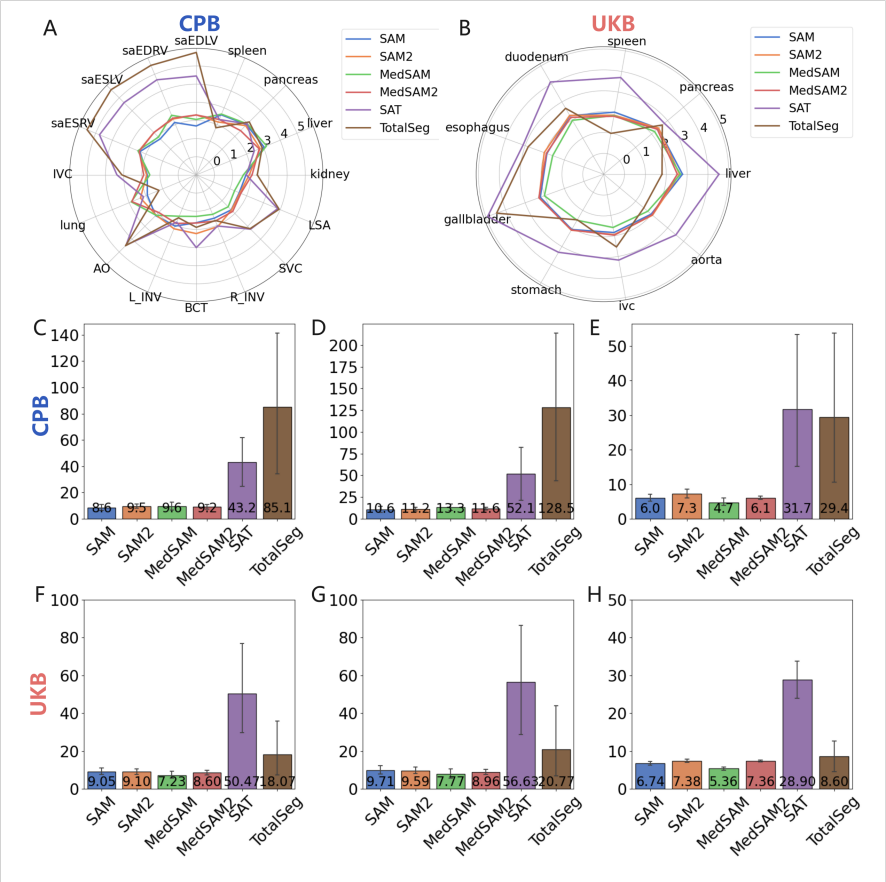

Figure S2: **Robustness of boundary alignment (HD95) across diverse models and anatomical targets. A, B,** Per-structure performance radar charts: Comparison of HD95 metrics across individual organs and vessels in the **(A)** CPB and **(B)** UKB cohorts. Bbox-prompted foundation models exhibit superior edge-tracing capabilities on prominent organs but degrade on vascular trees, while alternative paradigms show heightened vulnerability on cardiac and specialized digestive targets. **C–H,** Aggregated performance summaries: Summary of mean HD95 scores across all targets, large organs, and vessels for the CPB **(C–E)** and UKB **(F–H)** cohorts, quantifying the geometric accuracy penalties incurred during fine-grained vascular segmentation.

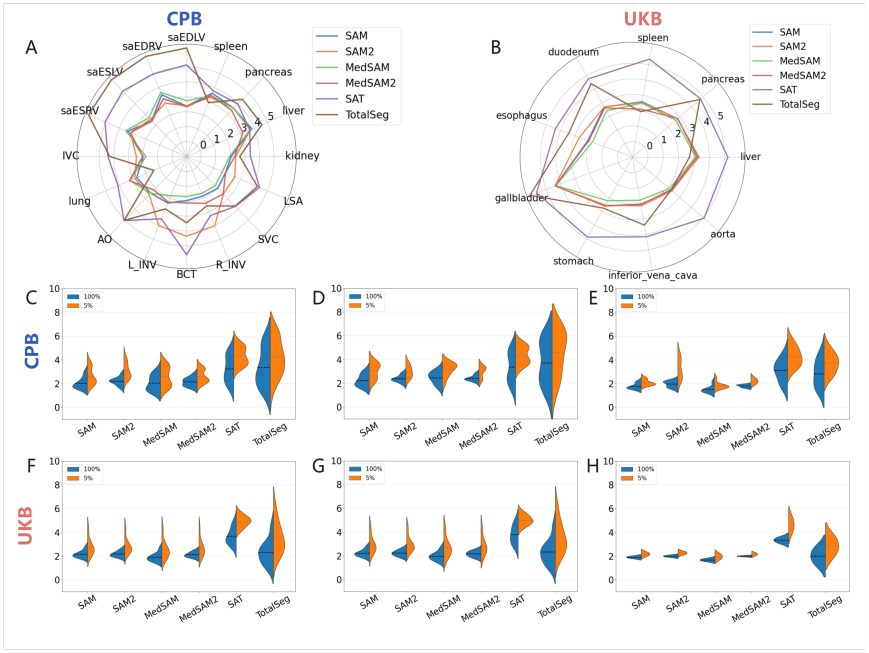

Figure S3: **Robustness and distribution shifts of boundary distance (HD95) in challenging clinical cases.** **A,B,** Per-structure performance under stress: Radar charts comparing the HD95 metrics on the worst 5% of cases for individual organs and vessels in the **(A)** CPB and **(B)** UKB datasets. Bbox-prompted SFMs exhibit superior resilience, whereas alternative paradigms display a severe performance collapse on complex target geometries. **C–H**, Performance distribution comparisons: Split-violin plots showcasing the HD95 variations between the worst-case 5% cohort (orange) and the entire population (blue) across all targets, organs, and vessels for CPB **(C–E)** and UKB **(F–H)** cohorts. The marked upward migration and elongation of the orange violins for SAT and TotalSeg highlight their systematic lack of robustness under challenging clinical scenarios.

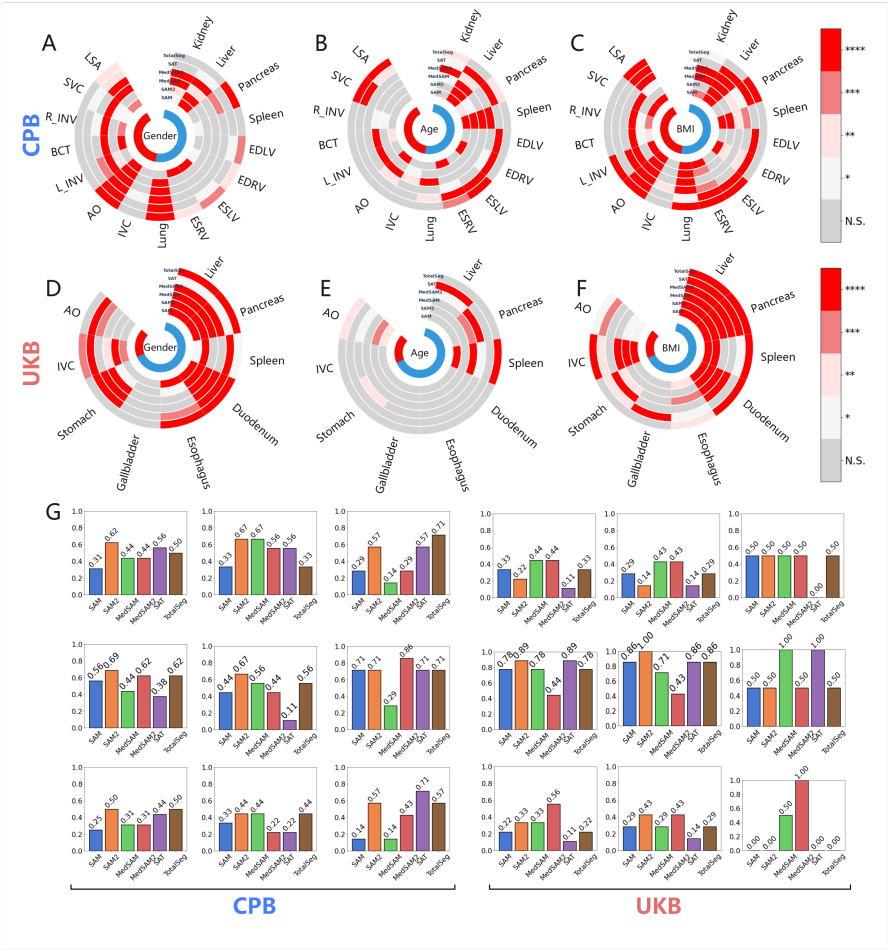

Figure S4: **Evaluation of demographic equity across foundation segmentation models based on HD95. A–F,** Mapping of algorithmic performance bias: Circular charts showing the statistical significance of performance disparities across demographic strata in the CPB **(A–C)** and UKB **(D–F)** datasets for gender, age, and BMI. Red regions indicate significant inter-subgroup differences in boundary distance errors detected by one-way ANOVA. **G,** Subgroup-specific equity profiles: Detailed bar plots compiling the overall fairness scores (the percentage of organs with $p > 0.05$) across diverse structural and demographic categories in both cohorts. The distributions illustrate the pervasive trade-offs between model accuracy and population equity across independent imaging resources. Statistical significance: $*p \leq 0.05$; $**p \leq 0.01$; $***p \leq 0.001$; $****p \leq 0.0001$; N.S., not significant.

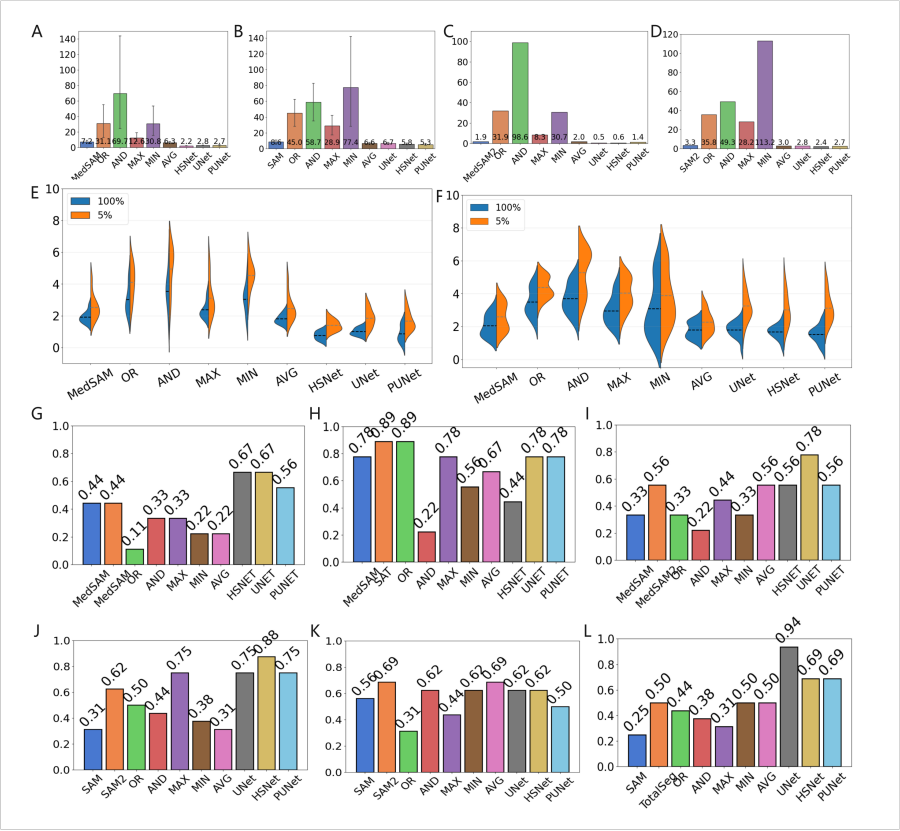

Figure S5: **Evaluation of architectural synergy and multi-source distillation based on HD95. A–D,** Aggregated boundary accuracy and inter-organ variance: Summaries of overall HD95 scores and standard deviations across the CPB **(A, C)** and UKB **(B, D)** datasets. Model synergy tracks successfully establish the lowest boundary placement errors and performance volatility compared to standalone SFMs. **E, F,** Robustness distribution profiles: Split-violin plots comparing HD95 dynamics between the entire cohort (blue) and the worst-case 5% sub-population (orange) for CPB **(E)** and UKB **(F)**. Synergistic methods effectively insulate the model from extreme anatomical distortions and boundary collapses. **G–L**, Subgroup equity analysis: Detailed fairness scores (the percentage of organs with p ¿ 0.05) stratified by gender, age, and BMI attributes. The collective frameworks retain high demographic impartiality without sacrificing global segmentation performance boundaries.

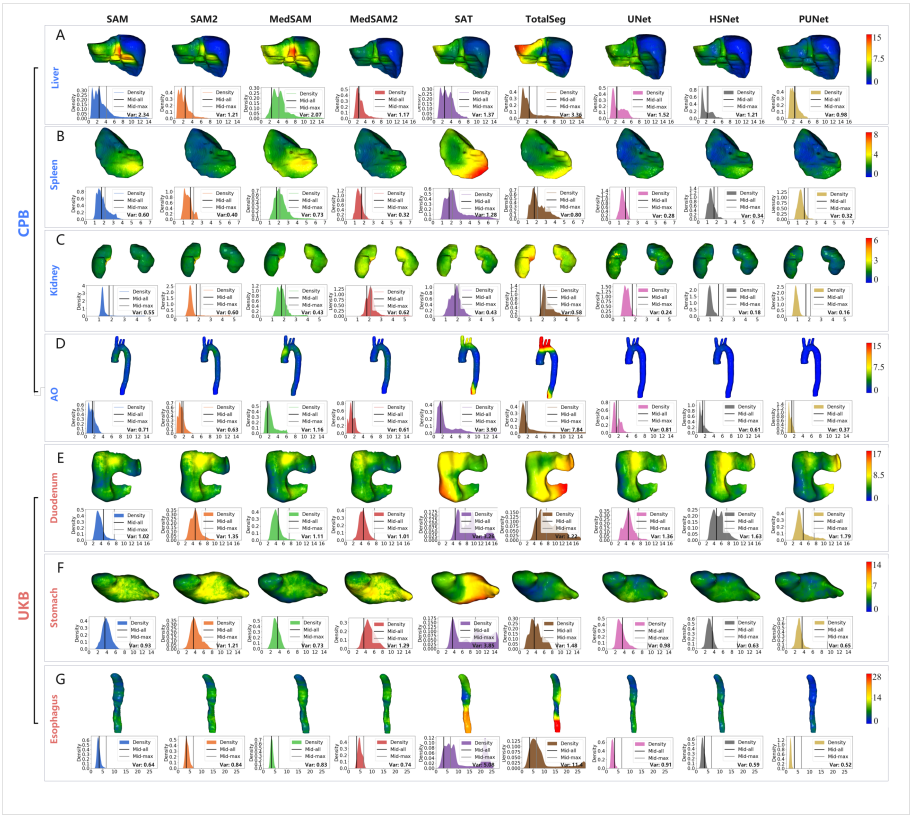

Figure S6: **Multi-source knowledge distillation homogenizing and mitigating the spatial divergence of anatomical error distributions. A–D,** Population-averaged spatial error topographies within the CPB cohort: 3D structural mapping and corresponding voxel-level probability density profiles for the **(A)** liver, **(B)** spleen, **(C)** kidney, and **(D)** aorta from the CPB data distribution. **E–G,** Population-averaged spatial error topographies within the external UKB cohort: 3D structural mapping and corresponding voxel-level probability density profiles for the **(E)** duodenum, **(F)** stomach, and **(G)** esophagus from the UKB data distribution. Color gradients projected onto the 3D anatomical atlases represent the mean Euclidean distance to the expert-curated manual ground-truth segmentations, where red spectrums denote regions of localized geometric discrepancy and blue spectrums denote high-fidelity alignment. Note that while standalone monolithic architectures present paradigm-specific, volatile error patterns (e.g., severe boundary collapses in SAT and TotalSeg) along with highly inflated spatial variance (Var), the distilled student frameworks (UNet, HSNet, and particularly PUNet) successfully enforce a smoother spatial distribution, dramatically compressing the overall error magnitudes and minimizing regional error variance across complex cross-cohort targets.