Expectation-Maximization as the Engine of Scalable Medical Intelligence

Wenxuan Li[1] Pedro R. A. S. Bassi[1,2,3] Tianyu Lin[1] Yu-Cheng Chou[1] Jakob Wasserthal[4]
Xinze Zhou[1] Qi Chen[1] Fabian Isensee[5] Yannick Kirchhoff[5] Maximilian Rokuss[5]
Saikat Roy[5] Constantin Ulrich[5] Klaus Maier-Hein[5] Szymon Płotka[6] Xiaoxi Chen[7]
Kang Wang[8] Yang Yang[8] Daguang Xu[9] Kai Ding[10] Yucheng Tang[9]
Alan L. Yuille[1] Zongwei Zhou[1,*]

[1]Johns Hopkins University [2]University of Bologna [3]Italian Institute of Technology
[4]University Hospital Basel [5]DKFZ [6]Jagiellonian University
[7]University of Illinois Urbana–Champaign [8]University of California, San Francisco
[9]NVIDIA [10]Johns Hopkins Medicine

Code, Dataset, and Models: https://github.com/MrGiovanni/ScaleMAI

Abstract

*Large, high-quality, annotated datasets are the foundation of medical AI research, but constructing even a small, moderate-quality, annotated dataset can take years of effort from multidisciplinary teams. Although active learning can prioritize what to annotate, scaling up still requires extensive manual efforts to revise the noisy annotations. We formulate this as a missing-data problem and develop ScaleMAI, a framework that unifies data annotation and model development co-evolution through an Expectation–Maximization (EM) process. In this iterative process, the AI model automatically identifies and corrects the mistakes in annotations (**Expectation**), while the refined annotated data retrain the model to improve accuracy (**Maximization**). In addition to the classical EM algorithm, ScaleMAI brings human experts into the loop to review annotations that cannot be adequately addressed by either Expectation or Maximization step (<5%).*

*As a result, ScaleMAI progressively creates an annotated dataset of **47,315** CT scans (**4.8**× larger than the largest public dataset, PanTS [44]) including **4,163,720** per-voxel annotations for benign/malignant tumors and 88 anatomical structures. ScaleMAI iteratively trains a model that exceeds human expert performance in tumor diagnosis (**+7%**), and outperforms models developed from smaller, moderate-quality datasets, with statistically significant gains in tumor detection (**+10%**) and segmentation (**+14%**) on two prestigious benchmarks.*

1. Introduction

Medical artificial intelligence (AI) is showing striking potential to detect diseases, guide treatments, and improve patient outcomes [10, 16, 55, 72]. To build reliable AI models, researchers need large, high-quality, annotated datasets that capture the complexity of human anatomy and pathology [40, 42, 56]. Creating such datasets, however, is painstakingly slow. Each voxel-wise annotation requires careful review by human experts and often years of coordination across institutions. As a result, annotation quantity and quality have not kept pace with the rapid advance of AI models, leaving a widening gap between what algorithms can do and the data they have to learn from.

This gap points to a deeper issue: data annotation and model development are still treated as two independent efforts [6, 43]. Models rely on the quality of data, but they rarely help improve it. Annotation noises persist through training, while model advances remain constrained by limited data scale. To break this cycle, we ask an important question: *Can a model help build the dataset that trains it?* In this view, the model is both a learner and an editor of its own training data.

The idea is inspired by Expectation-Maximization (EM) [17], a classic algorithm for refining estimates when ground truth is only partly known. We reimagine that principle for medical AI. Our approach begins with a few publicly available datasets that have incomplete annotations (listed in Figure 4). First, a model reviews existing annotations and flags uncertain or inconsistent regions (*Expectation*). Human ex-

*Correspondence to Zongwei Zhou (ZZHOU82@JH.EDU)
This is an extension of our previous manuscript, titled *"ScaleMAI: Accelerating the development of trusted datasets and AI models."*

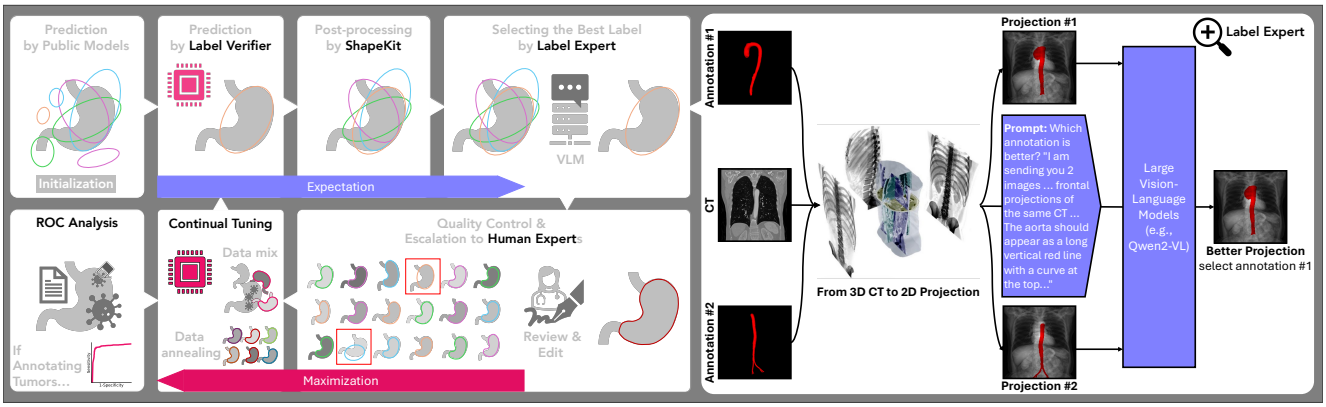

Figure 1. ScaleMAI reimagines the classic Expectation–Maximization (EM) algorithm [17] for the problem of building large, high-quality medical datasets when expert annotations are scarce and noisy. Instead of training a model on a fixed dataset and stopping there, we let the model and the dataset improve each other in a loop. At a high level, the model first "overfits" to the current dataset and then acts as a critic of that same dataset: wherever its predictions and the existing annotations disagree strongly, we treat this as missing or unreliable information. In the ***Expectation*** step, automatic tools (Label Verifier and Label Expert) use this disagreement to correct easy annotation errors and highlight only the most doubtful regions for human review. In the ***Maximization*** step, human experts focus on those few flagged cases, refine the annotations with the help of ROC-guided prioritization, and the model is retrained on this improved dataset using a mixture of unlabeled, synthetic, and selectively sampled scans. Repeating this cycle gradually turns a small, imperfect dataset into a large, expert-level resource, while keeping human effort concentrated on the <5% of annotations where the AI remains uncertain.

perts then correct only those areas, and the improved data retrain the model (*Maximization*). Repeating this cycle allows both the dataset and the model to improve together, gradually converging toward expert-level annotation quality. Unlike classical EM process, we integrate selective human feedback when both E and M steps fail to resolve annotation noise.

We call this **ScaleMAI**, a framework that connects data and model development through an iterative loop inspired by the EM process (§2). This design transforms model training from a passive process into a self-reinforcing process. The AI no longer consumes data blindly—it audits, questions, and learns from it. Human experts step in only when the model is uncertain, focusing on the small portion of annotations that need direct attention (<5% of the dataset). This targeted workflow reduces years of work to a few months of focused refinement. The process continues until both the annotations and the model achieve a level of performance comparable to human experts.

We applied this ScaleMAI framework to abdominal CT imaging, making two contributions.

1. **An open, expert-level annotated dataset** (§3), comprising 47,315 CT scans with precise per-voxel annotations of benign and malignant pancreatic tumors, along with 88 surrounding structures. Sourced from 112 hospitals, this dataset includes imaging metadata such as patient sex, age, contrast phase, diagnosis, spacing, and scanner details, and also includes structured and narrative radiology reports.
2. **An open, expert-level performing model** (§4), developed through progressive EM process, can exceed human expert performance in tumor diagnosis (+7%) and significantly outperforms models trained on smaller, moderate-quality datasets, achieving notable gains in detection (+10%), and segmentation (+14%) across two prestigious tumor benchmarks.

2. ScaleMAI

2.1. The Expectation Step

2.1.1. Label Verifier

Intuition. We assume that if a segmentation model can be trained on a large dataset and then reproduce the existing annotations on held-out scans, those annotations are at least self-consistent. In contrast, when the model strongly disagrees with an annotation, it is a signal that something is missing or structurally wrong in that annotation.

Mechanism. Label Verifier is an automatic quality controller in the *Expectation step* that identifies and refines noisy annotations. We train an nnU-Net [32] ($\mathcal{M}_{\text{verifier}}$) on the incompletely annotated dataset $\mathcal{D}_{\text{PanTS-XL-pseudo}}$ and evaluate it on the same dataset to assess annotation reliability. For each anatomical structure, we compute the Dice Similarity Coefficient (DSC) between $\mathcal{M}_{\text{verifier}}$'s prediction and the existing pseudo annotation in $\mathcal{D}_{\text{PanTS-XL-pseudo}}$. High DSC indicates self-consistency with $\mathcal{M}_{\text{verifier}}$ learning, while low DSC suggests potential noise or structural inconsistency, such as missing annotated regions or mismatches introduced by different annotation protocols.

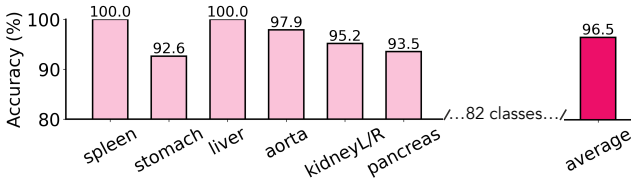

Figure 2. **Label Expert selects higher-quality annotations across diverse anatomical structures.** Evaluated on a 3,000 CT scan validation set, Label Expert achieves 96.5% accuracy across all 88 classes. We report results on organ-at-risk for pancreatic tumors. Label Expert consistently chooses the better annotation, including challenging cases such as the pancreas, where it correctly selected 116 of 124 comparisons (93.5% accuracy). This indicates its effectiveness in identifying higher-quality labels.

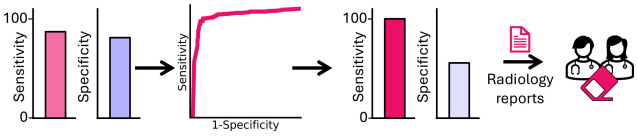

Figure 3. **ROC analysis for pancreatic tumor annotation.** Annotating per-voxel tumors is time-consuming. Our ROC analysis strategy biases AI predictions toward high sensitivity. Inevitably, this generates more false positives, but removing them is much faster and easier (<5 sec/tumor) than creating annotations from scratch (4–5 min/tumor). False positives in non-tumor CT scans can be automatically removed using radiology reports, and false positives in tumor CT scans can be erased with a few clicks. We achieved 99% sensitivity for pancreatic tumor detection with only 0.6 false positives per scan—reducing annotation time by up to 92% comparing to traditional methods.

Update rule. To avoid over-trusting $\mathcal{M}_{\text{verifier}}$, we only perform automatic replacement in the extreme case $\text{DSC} = 0$, where model and annotation have no spatial overlap (either one is empty and the other is non-empty, or both are in disjoint locations). In these cases, at least one of the two must be wrong, and our validation below shows the model prediction is almost always preferred by human experts. For all other DSC values, Label Verifier does not overwrite the annotation but simply forwards the case to the next stage (Label Expert or human review).

Validation. We have Label Verifier detect and refine 12,000 noisy pseudo annotations. From a random sample of 600 cases, two human experts confirmed all 600 as true errors and preferred the replacements from $\mathcal{M}_{\text{verifier}}$ over the original pseudo annotations. Label Verifier refined an average of 35.6% of noisy pseudo annotations (Appendix Table 3).

2.1.2. Label Expert

Intuition. A majority of annotation errors are obvious even to non-experts and can be detected by vision-language models (VLMs) [69] trained on diverse and extensive image-text datasets, given their strong performance in various image understanding tasks [2, 35, 39, 45, 57]. Examples of such obvious errors include organ misplacement, abnormal shapes, disconnections, multiple predictions for a single organ, noise artifacts, and annotation inconsistencies due to poor CT quality.

Mechanism. We propose Label Expert to select the higher quality annotations using a VLM guided by anatomical knowledge. At the initialization step for EM process, Label Expert compares model predictions generated by 19 existing models from the Touchstone Benchmark [5], with the existing pseudo annotations in $\mathcal{D}_{\text{PanTS-XL-pseudo}}$. During the EM process loop, Label Expert compares predictions with low DSC (<50%) from $\mathcal{M}_{\text{verifier}}$ (§2.1.1) with the existing pseudo annotations in $\mathcal{D}_{\text{PanTS-XL-pseudo}}$. Since pre-existing VLMs are trained on 2D natural images, they cannot directly analyze 3D CT scans. To address this, we project 3D CT scans and labels into 2D images, using a front-view projection. These projections resemble 2D X-rays with overlaid labels in red, as shown in Figure 1. The VLM then evaluates these projections with prompts designed to guide its decision-making. We use the aorta as an example. The prompt teaches the VLM that '*aorta should appear as a long vertical red line with a curve at the top.*' When comparing two annotations, the VLM determines that annotation #1 matches the description better than annotation #2.

Selection strategy. We use a lightweight tournament scheme: all pseudo annotations start from ShapeKit [49] post-processing; the VLM then compare each candidate annotation sequentially and in each pairwise comparison, we keep whichever annotation the VLM prefers. After all comparisons, the remaining annotation is taken as the updated pseudo annotation.

Validation. We evaluate Label Expert on a held-out validation set of 3,000 CT scans, each providing expert-created per-voxel annotations for all anatomical structures present in the scan. Across 88 classes, Label Expert achieves 96.5% accuracy in identifying the better annotation. For organ-at-risk structures relevant to pancreatic tumors (Figure 2), it also performs reliably: it selects the better pancreas annotation in 116 of 124 comparisons (93.5%) and the better aorta annotation in 1,674 of 1,710 comparisons (97.9%). By automating 34.7 million pairwise comparisons, Label Expert reduces human review to under 5%, while traditional error-detection methods miss roughly 80% of such errors [8, 67].

2.2. The Maximization Step

2.2.1. ROC Analysis

Intuition. For pancreatic tumor annotation, missing a tumor is far more costly than marking a few extra false positives. We observed that creating a new tumor mask from scratch takes 4–5 minutes, while removing an AI-generated false positive takes only a few seconds. This motivates an an-

notation strategy that prioritizes high sensitivity, even if it produces additional false positives that are easy to clean up.

Mechanism. We propose an efficient strategy, called *ROC analysis*, to assist human experts in annotating tumors within a large-scale dataset (e.g., 47,315 CT scans in PanTS-XL). During the EM loop, the improving models produce increasingly accurate tumor predictions. We analyze its receiver operating characteristic (ROC) curve to select a prediction threshold that achieves near-perfect sensitivity while keeping false positives manageable, as shown in Figure 3. (1) False positives in non-tumor CTs are automatically removed using radiology reports. (2) False positives in tumor CTs are quickly corrected using open-source tools [11] that allow radiologists to erase small regions with a few clicks (<5 seconds).

Validation. This sensitivity-first strategy dramatically reduces expert workload. The approach achieves 99% sensitivity with 0.6 false positives per scan, reducing annotation time by up to 92% compared to annotating every tumor voxel from scratch.

2.2.2. Continual Tuning

Mechanism. To optimize the training of Flagship Model, we incorporate a combination of data mix and data annealing strategies. ***Data mix*** consists of three primary data types. *First*, unlabeled data supports self-supervised representation learning. This approach leverages the large amount of raw clinical CT scans produced daily, requiring no manual annotations. The learned representations effectively regularize Flagship Model, enabling faster and more efficient learning of segmentation tasks with reduced reliance on annotated data [77, 79]. *Second*, synthetic data introduces variations that may not be fully represented in the training dataset, such as differences in patient demographics, scanner types, contrast phases, or tumor characteristics (e.g., location, shape, texture, size, intensity) [19, 28–30]. This diversity helps the Flagship Model adapt better to out-of-distribution cases, improving its generalization. *Third*, selective data targets the most challenging regions of CT scans as identified by loss function during training [12, 15, 75, 76, 78]. By prioritizing repeated sampling of these regions, we can avoid Flagship Model learning from non-informative areas such as air, bedding, or irrelevant anatomical regions. This ensures that Flagship Model focuses on clinically relevant areas, such as the pancreas or abdominal region. Finally, we apply ***data annealing*** [20] by fine-tuning Flagship Model on a subset of scans with expert-created per-voxel annotations for all classes.

Validation. We evaluate Flagship Model on tumor diagnosis, detection, and segmentation as detailed in §4.

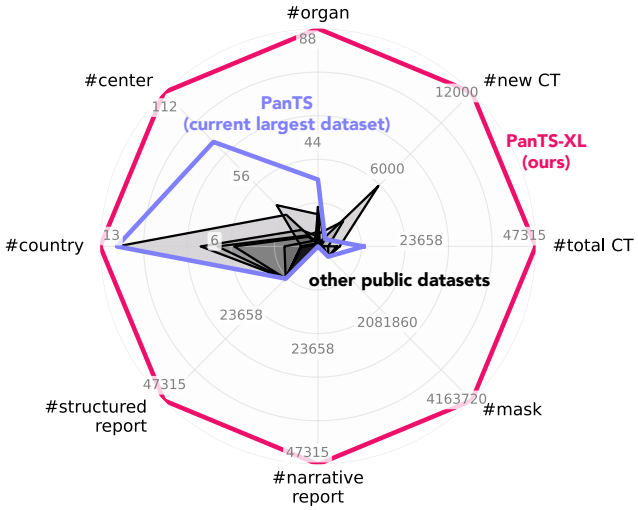

Figure 4. **Comparison of pancreatic and abdominal CT datasets.** We compare PanTS-XL with public datasets along eight axes: number of CT scans, first-time public scans, annotated structures, contributing centers, contributing countries, availability of structured and narrative reports, and total per-voxel annotations. *Earlier pancreatic and abdominal datasets* were already benchmarked in the PanTS study [44]; therefore, our comparison focuses on *PanTS* and *PanTS-XL*. A detailed comparison with public available datasets is provided in Appendix B.3.

2.3. An Executable Summary

At each EM loop, every annotation passes through the following decision flow:

- *Label Verifier:* If the model ($\mathcal{M}_{\text{verifier}}$) prediction and the current pseudo annotation have no overlap (DSC=0), we automatically replace the annotation with the $\mathcal{M}_{\text{verifier}}$ prediction. This captures clear structural errors.
- *Label Expert:* For remaining cases with low but non-zero consistency (DSC<0.5), we call Label Expert and it compares the candidate annotations and selects the better one.
- *Human escalation:* If Label Expert still cannot confidently resolve a case (e.g., conflicting VLM signals, extremely unusual anatomy, or low model confidence across all candidates), the case is escalated to human experts in the Maximization step. In practice, less than 5% of annotations follow this path.

Thresholds (DSC=0 for automatic replacement and DSC<0.5 for VLM review) were chosen empirically to maximize precision on the held-out validation set (3,000 CT scans) and remain fixed throughout all experiments.

3. Contribution #1: PanTS-XL Dataset

3.1. Dataset Overview

We construct PanTS-XL, a large-scale CT dataset comprising 47,315 scans with per-voxel annotations for pancreatic

tumors and 88 surrounding anatomical structures, sourced from 112 centers across 13 countries. To ensure annotation consistency, we developed a rigorous standard grounded in human sectional anatomy [18] and used it to guide human experts throughout the refinement process (Appendix B.1). In addition to annotations, PanTS-XL includes 12,000 first-time public CT scans, paired structured and narrative radiology reports, and comprehensive imaging metadata, including scanner types, contrast phases, and patient demographics (age and sex). Figure 4 compares PanTS-XL with prior CT datasets along several axes, including number of scans, annotated structures, contributing centers and countries, availability of clinical reports, and total per-voxel annotations. Because earlier public tumor and organ datasets were already benchmarked in the PanTS study [44], we focus our discussion on the comparison with PanTS, the previous largest one. Relative to PanTS, PanTS-XL (1) contributes 12,000 newly released CT scans, (2) expands the annotated classes from 27 to 88, (3) increases total per-voxel annotations from 277,228 to 4,163,720 (15×), and (4) increases source diversity from 76 to 112 centers across 13 countries. *We will make PanTS-XL publicly available.*

3.2. Gold Standard vs. Silver Standard Annotation

We distinguish two annotation types. *(1) Gold standard annotations* are created by human experts and verified by pathology, providing highly accurate labels but available only in limited quantity due to the clinical verification workflow. *(2) Silver standard annotations* are created using imaging alone, without pathology confirmation, enabling far larger datasets. Gold standard annotations are essential for assessing clinical correctness, while silver standard datasets increase the scale and diversity needed to train high-performing models (Table 2). PanTS-XL and Flagship Model improve together through ScaleMAI. As demonstrated in the reader study (§3.2.1) and anatomical structure evaluation (§3.2.2), Flagship Model eventually reaches human-expert–level performance in tumor detection and diagnosis (Figure 5), and produces anatomical structure annotations whose quality is validated on an independent, manually annotated test set (Table 1). Once Flagship Model achieves this level of accuracy, further human refinement provides little additional benefit. We therefore stop ScaleMAI and output $\mathcal{D}_{\text{PanTS-XL-pseudo}}$ as PanTS-XL, treating it as a silver-standard dataset produced by a model whose performance matches that of trained human experts.

3.2.1. Reader Study: Tumor Detection & Diagnosis

We conducted a multi-institution, multi-reader study to compare Flagship Model with human readers of varying experience levels on *tumor detection* and *diagnosis*.

Settings. Thirteen board-certified human readers participated: 6 juniors (<8 years), 5 seniors (8–15 years), and 2 experts (>15 years). Each reader independently reviewed contrast-enhanced abdominal CT scans from 50 patients (100 scans), covering normal cases and three pancreatic tumor subtypes (cysts, PDAC, PNET). Readers were blinded to case distribution and performed both tumor localization (3D point marking) and subtype classification. Flagship Model was evaluated under identical settings on a larger cohort of 982 patients (1,964 scans).

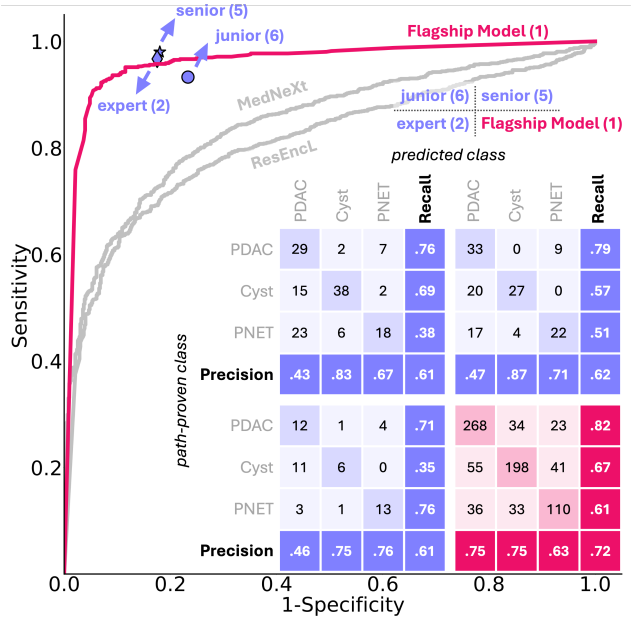

Figure 5. **Flagship Model matches human readers in *tumor detection* and surpasses them in *tumor diagnosis*.** We compare Flagship Model with 13 human readers (6 junior, 5 senior, 2 expert) on pancreatic tumor detection and diagnosis. Each reader independently evaluated 50 patients (100 contrast-enhanced CT scans); Flagship Model was evaluated on a larger cohort of 982 patients (1,964 scans). ***Tumor detection.*** ROC curves (top left) show that Flagship Model achieves an AUC of 0.961, surpassing MedNeXt [61] (0.846) by 13.5% and ResEncL [33] (0.810) by 15.1%, while matching the sensitivity–specificity performance of human readers. ***Tumor diagnosis.*** Confusion matrices (bottom right) show that Flagship Model attains 72% accuracy, outperforming junior (61%; +11%), senior (66%; +6%), and expert readers (69%; +3%) across PDAC, cyst, and PNET classification. Additional reader-study analyses are provided in §B.2.

Results. As shown in Fig. 5, Flagship Model achieves human-expert-level tumor detection performance, with an AUC of 0.961, providing a +13.5% gain over MedNeXt (0.846) and +15.1% over ResEncL (0.810). In tumor diagnosis, Flagship Model attains 72% accuracy, surpassing junior (+11%, 61%), senior (+6%, 66%), and expert readers (+3%, 69%). Class-wise results (PDAC, cyst, PNET) are detailed in §4.1. Additional analyses for size-stratified detection performance are provided in Appendix B.2.

Table 1. **Significance of PanTS-XL dataset.** AI models trained on PanTS-XL significantly outperform those trained on smaller datasets in terms of generalization ability. We compare the segmentation performance of nnU-Net [33] trained on BTCV, WORD, AbdomenAtlas 1.0, and PanTS-XL, evaluated on a manually annotated out-of-distribution proprietary dataset (N=300). The model trained on PanTS-XL achieved the highest DSC scores across all anatomical structures, demonstrating superior robustness and generalization. Performance is reported as median (Q1–Q3) of DSC scores, where Q1 and Q3 denote the first and third quartiles. In addition, we have further performed a one-sided Wilcoxon signed rank test between the best-performing model and others [71]. The performance gain is statistically significant at the $P = 0.05$ level, with highlighting in a pink box.

			out-of-distribution test on the proprietary dataset (N=300)				
training dataset	# of CTs	annotators	spleen	kidneyR	kidneyL	gallbladder	liver
BTCV [38]	47	human	93.6 (75.6–95.4)	43.9 (0.9–89.9)	94.8 (93.0–95.5)	77.1 (30.5–88.3)	95.4 (94.7–95.9)
WORD [50]	120	human	93.0 (89.7–94.3)	95.5 (94.9–95.9)	95.1 (92.8–95.8)	78.5 (51.6–86.5)	94.8 (93.8–95.5)
AbdomenAtlas 1.0 [56]	5,195	human-AI	95.8 (95.1–96.5)	93.2 (91.9–94.4)	92.8 (91.3–93.9)	88.2 (82.0–90.9)	96.4 (95.8–96.9)
PanTS-XL	47,315	human-AI	**96.2** (95.2–96.9)	**97.7** (97.4–98.0)	**97.6** (97.3–97.9)	**88.5** (80.6–92.1)	**96.7** (96.2–97.2)
Δ			**+0.4**	**+2.2**	**+2.5**	**+0.3**	**+0.3**
			stomach	aorta	postcava	pancreas	average
BTCV [38]	47	human	92.0 (87.1–94.0)	61.3 (19.6–83.3)	69.1 (36.8–80.6)	74.5 (66.6–79.5)	72.5 (61.8–81.3)
WORD [50]	120	human	90.7 (87.6–92.6)	-	-	75.9 (68.2–80.9)	87.1 (80.8–89.8)
AbdomenAtlas 1.0 [56]	5,195	human-AI	94.7 (93.0–95.5)	90.4 (87.6–91.8)	81.2 (75.1–84.9)	82.9 (78.7–85.9)	89.8 (87.9–91.2)
PanTS-XL	47,315	human-AI	**95.8** (94.3–96.4)	**91.8** (88.2–94.4)	**85.8** (82.1–88.8)	**85.7** (81.8–88.1)	**92.0** (89.7–93.2)
Δ			**+1.1**	**+1.4**	**+4.6**	**+2.8**	**+2.2**

3.2.2. High Quality Anatomical Structure Annotation

We assess the quality of anatomical structure annotations produced through ScaleMAI by comparing nnU-Net models trained on four abdominal CT datasets, including BTCV [38], WORD [51], AbdomenAtlas 1.0 [56], and our PanTS-XL. All models were evaluated on the same manually annotated proprietary test set (N=300). We focused on nine organs-at-risk for pancreatic tumors, as these structures directly affect downstream tasks such as tumor staging and radiotherapy planning. As shown in Table 1, the model trained on PanTS-XL achieves the highest DSC across *all* evaluated structures, with consistent gains over models trained on smaller, manually annotated datasets. Median improvements range from +0.3% to +4.6%, indicating that anatomical structure annotations produced through ScaleMAI support stronger out-of-distribution generalization.

4. Contribution #2: Flagship Model

Baselines. We compare Flagship Model with three groups of publicly available baselines: (1) Swin UNETR [66], the top-performing model on the MSD leaderboard [4]; (2) DTI [46], the top-ranking model on the public PANORAMA benchmark [3]; and (3) top-performing models from the Touchstone benchmark [5], including MedNeXt [61], nnU-Net ResEncL [33], STU-Net-B [31], and UniSeg [73]. Swin UNETR is implemented within the MONAI framework [11], and other models use the self-configuring nnU-Net framework [32], providing standardized training and hyperparameter selection across datasets.

Evaluation metrics. We evaluated the performance of baselines and Flagship Model using standard metrics for tumor detection and segmentation (Table 2). For tumor detection, we report Sensitivity[1], Specificity, and F1-score. For tumor segmentation, we report Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD). Detailed definitions of all metrics are in Appendix C.1.

Datasets. We benchmark baselines and Flagship Model on three datasets: MSD-Pancreas [4], PANORAMA [3], and a proprietary dataset. All datasets contain per-voxel annotations for pancreatic tumors. MSD-Pancreas (N=281) is used to train public baselines, ensuring fair comparison. For out-of-distribution (OOD) evaluation, we use PANORAMA (N=1,964)[2] and the proprietary dataset (N=1,958). PANORAMA provides metadata including patient sex, age, scanner type, and tumor size. The proprietary dataset offers comparable metadata, with additional information on contrast phases (arterial/venous) and tumor subtypes. See Appendix C.2 for details of datasets attributes.

4.1. Tumor Diagnosis (+7% Accuracy)

We evaluate Flagship Model on three-class tumor diagnosis (PDAC, cyst, PNET)[3]. As shown in Figure 5, Flagship Model achieves an accuracy of 72%—an average +7% improvement over human readers, outperforming junior (61%; +11%), senior (66%; +6%), and expert readers (69%; +3%). For PDAC, Flagship Model attains a recall of 82%, ex-

[1]Sensitivity is evaluated at two levels: (1) *patient-wise sensitivity*, which measures whether a patient is correctly identified as having at least one tumor (Table 2); and (2) *tumor-wise sensitivity*, which measures whether each tumor instance is correctly detected based on intersection with ground-truth annotations (Table 6 in Appendix C.4).

[2]PANORAMA originally included 194 MSD-Pancreas cases and 80 NIH cases [59]; these are removed for a fair OOD assessment.

[3]Due to the absence of publicly available datasets suitable for benchmarking Flagship Model diagnosis performance, our evaluation compares Flagship Model exclusively with human reader performance.

Table 2. **Pancreatic tumor detection and segmentation performance.** We benchmark pre-existing models trained on MSD-Pancreas and PANORAMA against Flagship Model on PANORAMA (N=1,964) and a proprietary dataset (N=1,958). For detection, we report patient-wise sensitivity, specificity (proprietary only), and F1-score (proprietary only). For segmentation, we report median DSC and NSD with IQR. Best results per dataset are **bolded**. In addition, we have performed a one-sided Wilcoxon signed rank test between the best-performing model and others [71]. The performance gain is statistically significant at the $P = 0.05$ level, with highlighting in a pink box.

		PANORAMA (N=1,964)†			Proprietary (N=1,958)				
method	training set	Sens.	DSC	NSD	Sens.	Spec.	F1	DSC	NSD
Swin UNETR [66]	MSD-Pancreas	85.5 (494/578)	39.4 (9.5–64.3)	31.9 (13.3–52.2)	62.7 (837/1335)	16.7 (104/623)	59.1 (1674/2831)	11.4 (0.0–49.1)	11.2 (0.0–32.5)
UniSeg [73]	MSD-Pancreas	77.9 (450/578)	49.8 (1.1–72.1)	38.7 (6.5–64.2)	58.9 (786/1335)	78.5 (485/623)	65.9 (1572/2385)	12.2 (0.0–56.9)	9.1 (0.0–44.6)
ResEncL [33]	MSD-Pancreas	74.9 (433/578)	54.1 (0.0–72.2)	41.0 (1.8–66.4)	60.8 (812/1335)	87.0 (542/623)	68.0 (1624/2387)	22.6 (0.0–65.6)	13.4 (0.0–52.4)
STU-Net [31]	MSD-Pancreas	77.0 (445/578)	51.8 (0.6–73.1)	41.5 (4.4–67.2)	58.1 (775/1335)	84.4 (526/623)	66.7 (1550/2324)	13.5 (0.0–62.3)	11.2 (0.0–48.4)
MedNeXt [61]	MSD-Pancreas	73.5 (425/578)	54.4 (0.0–74.8)	40.7 (0.4–66.9)	63.7 (850/1335)	83.1 (518/623)	69.9 (1700/2431)	30.6 (0.0–69.9)	20.4 (0.0–60.1)
DTI [46]	PANORAMA	–	–	–	69.6 (929/1335)	88.1 (549/623)	79.6 (1935/2431)	42.7 (0.0–72.1)	35.7 (0.0–63.2)
Flagship Model	PanTS-XL	**88.1 (509/578)**	**56.8 (23.7–75.8)**	**44.5 (20.5–66.0)**	**86.2 (1151/1335)**	**88.3 (550/623)**	**84.9 (2302/2713)**	**68.6 (34.7–82.9)**	**63.3 (33.2–81.9)**
Δ		**+2.6**	**+2.4**	**+3.0**	**+16.6**	**+0.2**	**+5.3**	**+25.9**	**+27.6**

†PANORAMA annotates only PDAC, treating all other types of pancreatic tumors and healthy pancreases as *Normal*—specificity and F1-score cannot be computed.

ceeding human readers by 6-11%. For cysts, it reaches 67% recall, improving on expert readers (57%) by +10%. For PNET, where human recall ranges from 38% (junior) to 51% (senior/expert), Flagship Model achieves 61%, corresponding to +23% over juniors and +10% over senior/expert readers. Across all tumor types, Flagship Model consistently outperforms human readers, with the largest gains on PNET, the most challenging class.

4.2. Tumor Detection (+10% Sensitivity)

We evaluate pancreatic tumor detection on PANORAMA and a proprietary dataset (Table 2). Across both OOD benchmarks, Flagship Model attains the highest sensitivity, improving by +2.6% on PANORAMA and +16.6% on the proprietary dataset—an average gain of +10%. PANORAMA annotates only PDAC and treats all other tumors and healthy pancreases as *Normal*; therefore, only sensitivity is reported. On PANORAMA benchmark, Flagship Model attains 88.1% sensitivity, outperforming all MSD-trained models. This corresponds to a +2.6% improvement over the strongest model (Swin UNETR: 85.5%). Since DTI is trained on PANORAMA itself, its performance cannot be evaluated fairly on this dataset and is therefore omitted. On the larger and more diverse proprietary benchmark, Flagship Model outperforms both MSD-trained and PANORAMA-trained models. Flagship Model achieves 86.2% sensitivity, improving over the strongest MSD-trained models (MedNeXt: 63.7%) by +22.5%, and PANORAMA-trained model (DTI: 69.6%) by +16.6%. Flagship Model also achieves the highest specificity (88.3%; +0.2%) and F1-score (84.9%; +5.3%) over the strongest model on this dataset.

Metadata analysis. We analyze tumor-wise detection performance across patient demographics and acquisition variables on PANOROMA and the proprietary dataset. As shown in Figure 6, on PANORAMA, Flagship Model yields higher sensitivity than the top-performing model (MedNeXt) across almost all reported groups, with average gains of +17.1% across age, +31.7% across sex, and +10.5% across scanner manufacturers. On the proprietary dataset (see Appendix Figure 13), Flagship Model maintains its advantage under a broader set of variables. Improvements include +23.3% across age groups, +20.9% across sexes, +10.5% across races, +18.3% across tumor sub-types, +19.3% across tumor-sizes, and +19.3% across contrast phases. These results demonstrate that Flagship Model preserves strong detection performance even under substantial demographic and acquisition variability.

4.3. Tumor Segmentation (+14% DSC)

We evaluate tumor segmentation on PANORAMA and a proprietary dataset (Table 2). Flagship Model achieves the highest segmentation performance, improving DSC by +2.4% on PANORAMA and +25.9% on the proprietary dataset—an average gain of +14%. On PANORAMA benchmark, Flagship Model achieves the top segmentation performance, increasing DSC from 54.4% (MedNeXt) to 56.8% (+2.4) and NSD from 41.5% (STU-Net) to 44.5% (+3.0). On the proprietary one, Flagship Model outperforms both MSD-trained and PANORAMA-trained models. It reaches 68.6% DSC, improving over the strongest MSD-trained model (MedNeXt: 30.6%) by +38.0 and over DTI (42.7%) by +25.9. NSD shows a similar pattern: Flagship Model attains 63.3%, exceeding MedNeXt (20.4%) by +42.9 and DTI (35.7%) by +27.6.

Metadata analysis. On PANORAMA (Figure 6), Flagship Model achieves higher DSC than MedNeXt across all reported groups, with gains of +8.2% (age), +6.5% (sex), and +6.7% (scanner). On the proprietary dataset (Appendix Figure 13), Flagship Model again leads across all groups, with improvements of +27.9% (age), +22.9% (sex), +27.2% (race), +21.3% (tumor subtype), +23.8% (tumor size), and +21.3% (contrast phase). Overall, Flagship Model maintains strong segmentation performance across diverse de-

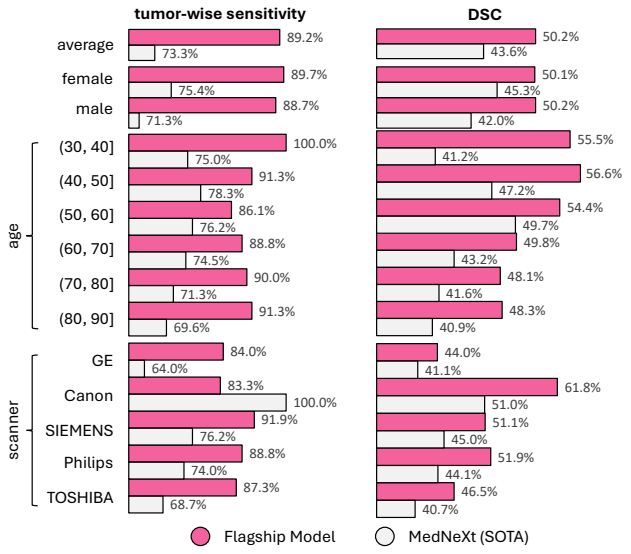

Figure 6. **Flagship Model achieves consistent gains across patient demographic and acquisition variables on the PANORAMA dataset.** Across all reported groups, Flagship Model improves over the top-performing model (MedNeXt) with average gains of +17.1% (age), +31.7% (sex), and +10.5% (scanner type) in sensitivity, and +8.2% (age), +6.5% (sex), and +6.7% (scanner manufacturer) in DSC. Full results for the proprietary dataset are provided in Appendix Figure 13.

mographic and acquisition variables.

5. Related Work

Previous studies treated data and model development separately. Most large datasets rely on manual annotation or crowd labeling, which cannot scale to complex 3D medical images. In contrast, we explore *whether a model help build the dataset that trains it*. ScaleMAI connects data and model improvement through an iterative process that automatically refines annotations, greatly reducing expert workload and enabling scalable medical AI.

Annotations. Recent large-scale natural image datasets, such as SA-1B [37], DataComp [21], Ego4D [22], and LAION-5B [63], were annotated through crowdsourcing, weak supervision, and prompt-based interactive labeling. These pipelines rely on large pools of annotators and simplified 2D primitives—points, boxes, or masks. However, such strategies do not transfer to 3D medical imaging, especially annotating tumors, where voxel-wise annotation demands anatomical accuracy, cross-slice consistency, and domain-specific knowledge. Given these challenges, recent created medical datasets speed up manual annotations through Active Learning [7, 36, 60, 64] and Human-in-the-Loop [27, 41, 53]. These approaches improve annotation efficiency by prioritizing uncertain samples for expert refine, yet they still require extensive human effort and cannot scale to larger datasets (tens of thousands of scans). To further reduce the manual annotation efforts, our EM process performs most annotation refinement (>95%) automatically, enabling efficient and scalable dataset annotations across tens of thousands of scans.

Datasets. Earlier abdominal and pancreatic CT datasets, including MSD-Pancreas [4], TCIA-Pancreas [59], PANORAMA [3], AMOS'22 [34], and FLARE'23 [54], have advanced cancer-related medical image segmentation but remain limited in scale, diversity, and clinical completeness. MSD-Pancreas includes only tumor (positive) cases, while TCIA-Pancreas contains only normal controls (negative). PANORAMA does not include truly normal controls because its so-called normal group excludes a type of pancreatic tumor but may include other types of pancreatic tumors. To address these limitations, PanTS [44] comprises 9,901 publicly available CT scans aggregated from multiple public datasets, with 28 annotated structures and detailed metadata, providing an important foundation for pancreatic tumor segmentation research. Building upon this progress, our PanTS-XL contributes 12,000 first-time public CT scans, sourced from 112 hospitals across 13 countries, provides paired structured and narrative radiology reports, and enriches each case with standardized clinical metadata. To the best of our knowledge, PanTS-XL is the *first* large-scale abdominal CT dataset that simultaneously provides images, segmentation masks, clinical metadata, and radiology reports, while also contributing over ten thousands newly collected CT scans. Moreover, PanTS-XL enables a wide range of clinical tasks beyond segmentation—spanning detection, diagnosis, and report-grounded multi-modal learning—and sets a new standard for comprehensive and scalable medical datasets.

Models. Recent progress in tumor segmentation has been driven by models such as nnU-Net [32, 33], C2FNAS [74], DiNTS [25], Swin-UNETR [66], Universal Model [47], and MedNeXt [61]. These methods have achieved strong results on the MSD benchmark [4] and related benchmarks through improved architecture design and training strategies. However, their performance remains limited because they depend on static datasets, where annotation quality and scale do not improve over time. Incomplete annotations and the lack of data diversity often prevent further progress, even with more advanced architectures [48]. We address this limitation by introducing an iterative EM process that continually refines both data (PanTS-XL) and model (Flagship Model) quality. The model retrains itself using improved annotations, gradually approaching expert-level accuracy.

6. Conclusion

ScaleMAI is directly motivated by the structure of the Ex-

pectation–Maximization (EM) algorithm. In classical EM, the data are incomplete and the true values are treated as latent variables to be estimated. We view large-scale medical datasets in the same way: the *missing data* are the unannotated or incorrectly annotated parts of each CT. In the *Expectation* step, ScaleMAI uses AI tools—Label Verifier and Expert—to propose improved annotations, filling in these missing or unreliable parts. In the *Maximization* step, the updated annotations are treated as completed data to retrain the model. Repeating this process improves both the model and the annotations, mirroring the iterative refinement and self-consistency behavior of EM. Unlike classical EM, ScaleMAI extends the E-step with VLM-based selection and selective human review, enabling data annotation at scale on real-world medical images, reducing manual efforts from years to months. ScaleMAI produces **PanTS-XL**, a large-scale dataset of 47,315 CT scans—4.8$\times$ larger than the prior largest one, PanTS [44]—with 4,163,720 per-voxel annotations for 88 classes across 112 centers. The resulting **Flagship Model** achieves best performance on two benchmarks, with significant gains of +7/+10/+14% in tumor diagnosis, detection, and segmentation, respectively.

Acknowledgments. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the National Institutes of Health (NIH) under Award Number R01EB037669. We would like to thank the Johns Hopkins Research IT team in IT@JH for their support and infrastructure resources where some of these analyses were conducted; especially DISCOVERY HPC. We thank Jaimie Patterson for writing a news article about this project. Paper content is covered by patents pending.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 7

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[3] N Alves, M Schuurmans, D Rutkowski, et al. The panorama study protocol: Pancreatic cancer diagnosis-radiologists meet ai. zenodo, 2024. 6, 8, 12, 14

[4] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):1–13, 2022. 6, 8, 12

[5] Pedro RAS Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Yannick Kirchhoff, Maximilian Rokuss, Ziyan Huang, Jin Ye, Junjun He, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus H. Maier-Hein, Paul Jaeger, Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Yong Xia, Zhaohu Xing, Lei Zhu, Yousef Sadegheih, Afshin Bozorgpour, Pratibha Kumari, Reza Azad, Dorit Merhof, Pengcheng Shi, Ting Ma, Yuxin Du, Fan Bai, Tiejun Huang, Bo Zhao, Haonan Wang, Xiaomeng Li, Hanxue Gu, Haoyu Dong, Jichen Yang, Maciej A. Mazurowski, Saumya Gupta, Linshan Wu, Jiaxin Zhuang, Hao Chen, Holger Roth, Daguang Xu, Matthew B. Blaschko, Sergio Decherchi, Andrea Cavalli, Alan L. Yuille, and Zongwei Zhou. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? *Conference on Neural Information Processing Systems*, 2024. 3, 6

[6] Pedro RAS Bassi, Mehmet Can Yavuz, Ibrahim Ethem Hamamci, Sezgin Er, Xiaoxi Chen, Wenxuan Li, Bjoern Menze, Sergio Decherchi, Andrea Cavalli, Kang Wang, Yang Yang, Alan Yuille, and Zongwei Zhou. Radgpt: Constructing 3d image-text tumor datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23720–23730, 2025. 1, 12

[7] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018. 8

[8] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37 (11):2514–2525, 2018. 3

[9] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019. 12

[10] Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature medicine*, 29(12): 3033–3043, 2023. 1

[11] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 4, 6

[12] Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu, Bihan Wen, Alan L Yuille, and Zongwei Zhou. Making your first choice: To address cold start problem in vision active learning. In *Medical Imaging with Deep Learning*. 2023. 4

[13] Luohai Chen, Wei Wang, Kaizhou Jin, Bing Yuan, Huangying Tan, Jian Sun, Yu Guo, Yanji Luo, Shi-Ting Feng, Xianjun Yu, et al. Special issue "the advance of solid tumor research in china": Prediction of sunitinib efficacy using computed tomography in patients with pancreatic neuroendocrine tumors. *International Journal of Cancer*, 152(1): 90–99, 2023. 12

[14] Qi Chen, Xinze Zhou, Chen Liu, Hao Chen, Wenxuan Li, Zekun Jiang, Ziyan Huang, Yuxuan Zhao, Dexin Yu, Junjun He, Yefeng Zheng, Ling Shao, Alan Yuille, and Zongwei Zhou. Scaling tumor segmentation: Best lessons from real and synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24001–24013, 2025. 12

[15] Yu-Cheng Chou, Zongwei Zhou, and Alan Yuille. Embracing massive medical data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–35. Springer, 2024. 4

[16] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24 (9):1342–1350, 2018. 1

[17] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977. 1, 2

[18] Adrian Kendal Dixon, David J Bowden, Bari M Logan, and Harold Ellis. *Human sectional anatomy: Pocket atlas of body sections, CT and MRI images*. CRC Press, 2017. 5

[19] Shiyi Du, Xiaosong Wang, Yongyi Lu, Yuyin Zhou, Shaoting Zhang, Alan Yuille, Kang Li, and Zongwei Zhou. Boosting dermatoscopic lesion segmentation via diffusion models with visual and textual prompts. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 4

[20] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 4, 7

[21] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023. 8

[22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 8

[23] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021. 14

[24] Xu Han, Jun Hong, Marsha Reyngold, Christopher Crane, John Cuaron, Carla Hajj, Justin Mann, Melissa Zinovoy, Hastings Greer, Ellen Yorke, et al. Deep-learning-based image registration and automatic segmentation of organs-at-risk in cone-beam ct scans from high-dose radiation treatment of pancreatic cancer. *Medical physics*, 48(6):3084–3095, 2021. 12

[25] Yufan He, Dong Yang, Holger Roth, Can Zhao, and Daguang Xu. Dints: Differentiable neural network topology search for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5841–5850, 2021. 8

[26] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019. 12

[27] Andreas Holzinger. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016. 8

[28] Qixin Hu, Junfei Xiao, Yixiong Chen, Shuwen Sun, Jie-Neng Chen, Alan Yuille, and Zongwei Zhou. Synthetic tumors make ai segment tumors better. *NeurIPS Workshop on Medical Imaging meets NeurIPS*, 2022. 4

[29] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou. Label-free liver tumor segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7422–7432, 2023.

[30] Qixin Hu, Alan Yuille, and Zongwei Zhou. Synthetic data as validation. *arXiv preprint arXiv:2310.16052*, 2023. 4

[31] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023. 6, 7, 16

[32] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 2, 6, 8

[33] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. *arXiv preprint arXiv:2404.09556*, 2024. 5, 6, 7, 8, 16

[34] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022. 8, 12

[35] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[36] Jun Kasai, Shota Sasaki, and Yasuhiro Mukaigawa. Reliability of active learning for medical image segmentation. *IEEE Access*, 9:160842–160854, 2021. 8

[37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 8

[38] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 6, 12

[39] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[40] Jianning Li, Zongwei Zhou, Jiancheng Yang, Antonio Pepe, Christina Gsaxner, Gijs Luijten, Chongyu Qu, Tiezheng Zhang, Xiaoxi Chen, Wenxuan Li, Yuan Jin, and Jan Egger. Medshapenet–a large-scale dataset of 3d medical shapes for computer vision. *Biomedical Engineering/Biomedizinische Technik*, (0), 2024. 1

[41] Wentao Li, Zhiqiang Chen, Hanchuan Zhang, and Dong Guo. Human-in-the-loop deep learning for medical image

segmentation: A review. *Pattern Recognition Letters*, 150: 61–71, 2021. 8

[42] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, Yutong Tang, Yining Cao, Haoqi Han, Zheyuan Zhang, Jiawei Liu, Tiezheng Zhang, Yujiu Ma, Jincheng Wang, Guang Zhang, Alan Yuille, and Zongwei Zhou. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, page 103285, 2024. 1

[43] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation? In *International Conference on Learning Representations*, 2024. 1, 12

[44] Wenxuan Li, Xinze Zhou, Qi Chen, Tianyu Lin, Pedro RAS Bassi, Szymon Plotka, Jaroslaw B Cwikla, Xiaoxi Chen, Chen Ye, Zheren Zhu, Yu-Cheng Chou, Kang Wang, Yucheng Tang, Alan L Yuille, and Zongwei Zhou. Pants: The pancreatic tumor segmentation dataset. *arXiv preprint arXiv:2507.01291*, 2025. 1, 4, 5, 8, 9, 12

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[46] Han Liu, Riqiang Gao, and Sasa Grbic. Ai-assisted early detection of pancreatic ductal adenocarcinoma on contrast-enhanced ct. *arXiv preprint arXiv:2503.10068*, 2025. 6, 7, 16

[47] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 8

[48] Jie Liu, Yixiao Zhang, Kang Wang, Mehmet Can Yavuz, Xiaoxi Chen, Yixuan Yuan, Haoliang Li, Yang Yang, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography. *Medical Image Analysis*, page 103226, 2024. 8

[49] Junqi Liu, Dongli He, Wenxuan Li, Ningyu Wang, Alan L Yuille, and Zongwei Zhou. Shapekit. In *International Workshop on Shape in Medical Imaging*, pages 44–58. Springer, 2025. 3

[50] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Tao Song, Xiaofan Zhang, Kang Li, Guotai Wang, and Shaoting Zhang. Word: Revisiting organs segmentation in the whole abdominal region. *arXiv preprint arXiv:2111.02403*, 2021. 6, 12

[51] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, page 102642, 2022. 6

[52] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 12

[53] Jun Ma, Yuyin Zhang, Yuhui Xu, Zongwei Fang, and Yefeng Chen. Human-in-the-loop medical image segmentation: Towards a unified framework. *Medical Image Analysis*, 87: 102791, 2023. 8

[54] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Ershuai Wang, Qin Zhou, Ziyan Huang, Pengju Lyu, Jian He, and Bo Wang. Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge. *arXiv preprint arXiv:2408.12534*, 2024. 8, 12

[55] NLST. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011. 1

[56] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou. Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. In *Conference on Neural Information Processing Systems*, 2023. 1, 6

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[58] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):1–9, 2020. 12

[59] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 556–564. Springer, 2015. 6, 8

[60] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Deep active learning for biomedical image segmentation. In *MICCAI*, pages 97–105. Springer, 2018. 8

[61] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 405–415. Springer, 2023. 5, 6, 7, 8, 16

[62] Jeffrey D Rudie, Hui-Ming Lin, Robyn L Ball, Sabeena Jalal, Luciano M Prevedello, Savvas Nicolaou, Brett S Marinelli, Adam E Flanders, Kirti Magudia, George Shih, et al. The rsna abdominal traumatic injury ct (ratic) dataset. *Radiology: Artificial Intelligence*, 6(6):e240101, 2024. 12

[63] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 8

[64] Burr Settles. Active learning literature survey. 2009. 8

[65] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016. 14

[66] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022. 6, 7, 8, 16

[67] Vanya V Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE transactions on medical imaging*, 36 (8):1597–1606, 2017. 3

[68] Vanya V Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 547–556. IEEE, 2018. 12

[69] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[70] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023. 12

[71] Manuel Wiesenfarth, Annika Reinke, Bennett A Landman, Matthias Eisenmann, Laura Aguilera Saiz, M Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. Methods and open-source toolkit for analyzing and visualizing challenge results. *Scientific reports*, 11(1):2369, 2021. 6, 7, 16

[72] Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, Alan L Yuille, Elliot K Fishman, and Ralph H Hruban. The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, 2022. 1

[73] Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 508–518. Springer, 2023. 6, 7, 16

[74] Qihang Yu, Dong Yang, Holger Roth, Yutong Bai, Yixiao Zhang, Alan L Yuille, and Daguang Xu. C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4126–4135, 2020. 8

[75] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7340–7351, 2017. 4

[76] Zongwei Zhou, Jae Shin, Ruibin Feng, R Todd Hurst, Christopher B Kendall, and Jianming Liang. Integrating active learning and transfer learning for carotid intima-media thickness video interpretation. *Journal of Digital Imaging*, 32(2):290–299, 2019. 4

[77] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 384–393. Springer, 2019. 4

[78] Zongwei Zhou, Jae Y Shin, Suryakanth R Gurudu, Michael B Gotway, and Jianming Liang. Active, continual fine tuning of convolutional neural networks for reducing annotation efforts. *Medical Image Analysis*, 71:101997, 2021. 4

[79] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021. 4

Appendix

Table of Contents

A. Technical Details of ScaleMAI

A.1. The Expectation Step

A.1.1. Label Verifier

Label Verifier serves as an automatic quality controller within the Expectation step of ScaleMAI. Its role is to identify annotations that are inconsistent with the statistical patterns learned by a segmentation model $\mathcal{M}_{\text{verifier}}$ trained on the current pseudo-annotations dataset $\mathcal{D}_{\text{PanTS-XL-pseudo}}$. The key intuition is that if $\mathcal{M}_{\text{verifier}}$ can accurately reproduce an annotation on held-out scans, that annotation is likely self-consistent; conversely, strong disagreement indicates potential noise or structural mistakes. By comparing $\mathcal{M}_{\text{verifier}}$ predictions with existing pseudo annotations for each structure, Label Verifier provides a principled mechanism to flag and refine erroneous annotations before they propagate into the next iteration of ScaleMAI.

Algorithm 1 Label Verifier: Model-Guided Detection & Refinement of Noisy Annotations

Input: Pseudo-annotated dataset $\mathcal{D}_{\text{PanTS-XL-pseudo}}$; trained verifier model $\mathcal{M}_{\text{verifier}}$ (nnU-Net trained on $\mathcal{D}_{\text{PanTS-XL-pseudo}}$); anatomical structure set $\mathcal{S}$.
Output: Updated pseudo-annotations $\hat{\mathcal{D}}_{\text{PanTS-XL-pseudo}}$ (with optional replacements when DSC = 0)
// Step 1: Compute agreement between model and annotation
for each sample (x, y_{pseudo}) in $\mathcal{D}_{\text{PanTS-XL-pseudo}}$ **do**
  Predict verifier mask: $y_{\text{verifier}} \leftarrow \mathcal{M}_{\text{verifier}}(x)$
  for each structure $s \in \mathcal{S}$ **do**
    Extract structure-specific masks:
      $y^s_{\text{pseudo}} \leftarrow$ mask of s in y_{pseudo}
      $y^s_{\text{verifier}} \leftarrow$ mask of s in y_{verifier}
    Compute Dice score:
      $\text{DSC}_s = \text{DSC}(y^s_{\text{pseudo}}, y^s_{\text{verifier}})$
  end for
  // Step 2: Apply update rule (only in extreme disagreement)
  for each structure $s \in \mathcal{S}$ **do**
    if $\text{DSC}_s = 0$ **then**
      $y^s_{\text{pseudo}} \leftarrow y^s_{\text{verifier}}$ {Replace annotation only when $\mathcal{M}_{\text{verifier}}$ prediction and pseudo annotation have no overlap}
    end if
  end for
  Store updated annotation y_{pseudo} in $\hat{\mathcal{D}}_{\text{PanTS-XL-pseudo}}$
end for
return $\hat{\mathcal{D}}_{\text{PanTS-XL-pseudo}}$

Label Verifier quantifies the agreement between the predictions of $\mathcal{M}_{\text{verifier}}$ and the existing annotations, and selectively replaces an annotation only when there is a complete mismatch ($\text{DSC} = 0$). To assess its practical impact, we measure how often Label Verifier detects and refines erroneous annotations during ScaleMAI's annotation of PanTS-XL. Table 3 summarizes representative anatomical structures with notable refinement rates, highlighting the effectiveness of Label Verifier—particularly for structures that are often absent in abdominal CT scans.

Table 3. **Label Verifier detects and refines 35.6% of annotation errors.** Label Verifier replaces a pseudo annotation only when it has no spatial overlap with $\mathcal{M}_{\text{verifier}}$ prediction (DSC = 0). In total, 51,454 erroneous annotations were refined iteratively through ScaleMAI. We report representative anatomical structures that show notable refinement rates. Label Verifier is particularly effective for some structures that are often absent in abdominal CT scans—such as the lung, prostate, rectum, and bladder—where existing annotations may contain false positives, resulting in no overlap with $\mathcal{M}_{\text{verifier}}$ prediction.

method	spleen	kidney	gallbladder	liver	stomach	aorta	pancreas	prostate	duodenum	femur	esophagus	lung	bladder	rectum	average
Label Verifier	28.0	24.6	48.8	8.3	38.4	10.0	14.0	52.8	40.8	48.7	24.6	47.4	49.6	49.4	35.6

A.1.2. Label Expert

Label Expert is designed to improve annotation quality by leveraging a VLM (e.g., Qwen2-VL [69]) with detailed anatomical knowledge as prompt. Given two candidate annotations for the same CT scan—one produced during initialization or by $\mathcal{M}_{\text{verifier}}$, and the other drawn from the existing pseudo-annotation set $\mathcal{D}_{\text{PanTS-XL-pseudo}}$ —Label Expert determines which annotation more faithfully reflects the underlying anatomy. Because existing VLMs operate on 2D natural images, rather than volumetric medical data, each 3D CT scan is first projected into a front-view image with the corresponding annotation overlaid in red. The VLM then evaluates these projections using structure-specific prompts that define the expected anatomical appearance, location, and continuity of each organ or structure, enabling an informed comparison between two annotations.

Algorithm 2 Label Expert: VLM-Guided Selection of Higher-Quality Annotation

Input: 3D CT scan x; pseudo-annotation $y_{\text{pseudo}} \in \mathcal{D}_{\text{PanTS-XL-pseudo}}$; candidate annotation y_{cand} (prediction from initialization or low-DSC output from $\mathcal{M}_{\text{verifier}}$ during ScaleMAI); anatomical structure set $\mathcal{S} = \{\text{Aorta, Postcava, Kidneys, Liver, Pancreas, Spleen, ... (88 classes)}\}$; VLM with structure-specific anatomical prompts.
Output: Selected annotation $\hat{y} \in \{y_{\text{cand}}, y_{\text{pseudo}}\}$

// Step 1: Prepare the two annotations for comparison
$Annotation_1 \leftarrow y_{\text{cand}}$
$Annotation_2 \leftarrow y_{\text{pseudo}}$

// Step 2: Generate 2D projections for VLM evaluation
$I_x \leftarrow$ front-view projection of CT scan x
$I_1 \leftarrow$ overlay $Annotation_1$ (red) on I_x
$I_2 \leftarrow$ overlay $Annotation_2$ (red) on I_x

// Step 3: Structure-wise comparison using VLM prompts
$score_1 \leftarrow 0$, $score_2 \leftarrow 0$
for each $structure \in \mathcal{S}$ **do**
    Construct anatomical prompt $p(structure)$
        // e.g., for the aorta: "the aorta appears as a long vertical red tube with a curve at the top"
    Query VLM with $(I_1, I_2, p(structure))$ to obtain preference $r_{structure} \in \{1, 2, \text{tie}\}$
    if $r_{structure} = 1$ **then**
        $score_1 \leftarrow score_1 + 1$
    else if $r_{structure} = 2$ **then**
        $score_2 \leftarrow score_2 + 1$
    end if
end for

// Step 4: Final selection
if $score_1 > score_2$ **then**
    $\hat{y} \leftarrow Annotation_1$ {VLM prefers candidate annotation}
else
    $\hat{y} \leftarrow Annotation_2$ {VLM prefers pseudo-annotation or tie}
end if
return $\hat{y}$

Label Expert selects the higher-quality annotation by scoring two candidate overlays across all anatomical structures with the assistance of a VLM. To carry out this procedure, each structure is assessed using a tailored prompt that describes its expected shape, location, and anatomical context. These prompts guide the VLM's reasoning when comparing the two projected annotations. The following prompt templates provide representative examples of the instructions given to the VLM for different structures, ensuring consistent and anatomically informed evaluation during the comparison process.

Label Expert Prompt: Aorta Annotation Comparison

You are given two images, Image1 and Image2, representing frontal projections of the same 3D CT scan and visually comparable to AP X-rays. Each image contains a red overlay representing the aorta, and the overlays differ between the two images.
Your task is to determine which image contains the more accurate aorta annotation.
Evaluation criteria:
1. *Cranial extension* — The aorta normally extends into the thoracic region; the better annotation shows a longer superior reach.
2. *Lumbar alignment* — If the lumbar spine is visible, the aorta overlay should descend to an anatomically plausible level.
3. *Shape and continuity* — The aorta should appear as a smooth tubular structure without fragmentation or abrupt widening.

After reviewing both images, conclude whether Image1 or Image2 contains the more anatomically accurate aorta annotation.

Label Expert Prompt: Postcava Annotation Comparison

You are given two images, Image1 and Image2, representing frontal projections of the same 3D CT scan and visually comparable to AP X-rays. Each image contains a red overlay representing the postcava, with different overlays shown in the two images.
Your task is to determine which overlay more accurately represents the postcava.
Evaluation criteria:
1. *Thoracic reach* — The postcava should extend into the thoracic region near the upper ribs.
2. *Lumbar descent* — If the lumbar spine is visible, the postcava overlay should descend into the upper lumbar region realistically.
3. *Continuity and form* — The postcava should appear as a continuous, narrow tubular structure without breaks or irregular deviations.

After evaluating both images, determine whether Image1 or Image2 provides the more anatomically correct postcava annotation.

Label Expert Prompt: Spleen Annotation Comparison

You are given two images, Image1 and Image2, representing frontal projections of the same 3D CT scan and visually comparable to AP X-rays. Each image contains a red overlay representing the spleen, and the overlays differ between the two images.
Your task is to determine which overlay more accurately represents the spleen.
Anatomical guidelines:
1. *Shape* — The spleen typically has an oval or crescent-like outline with smooth, continuous curvature.
2. *Smoothness* — A proper spleen contour should not contain irregular protrusions, abrupt angles, internal gaps, or jagged edges.
3. *Continuity* — The spleen is a single anatomical structure; the overlay should therefore form one continuous region.
4. *Location* — The spleen lies in the upper left abdomen (right side of the image in AP orientation), beneath the ribs and diaphragm and adjacent to the stomach and left kidney.

After reviewing both images, conclude whether Image1 or Image2 contains the more anatomically accurate spleen annotation.

Label Expert Prompt: Kidney Annotation Comparison

You are given two images, Image1 and Image2, representing frontal projections of the same 3D CT scan and visually comparable to AP X-rays. Each image contains red overlays representing the kidneys, which differ between the two images.

Your task is to determine which image contains the more accurate kidney annotation.
Anatomical guidelines:
1. *Number of structures* — Two kidneys should be present, appearing as two separate regions.
2. *Shape* — Kidneys have a bean-like outline with a concave side medially and a convex side laterally.
3. *Location* — Kidneys lie lateral to the spine near the lower ribs, typically at similar vertical levels.

After comparing both images, conclude whether Image1 or Image2 contains the more anatomically accurate kidney annotation.

Label Expert Prompt: Liver Annotation Comparison

You are given two images, Image1 and Image2, representing frontal projections of the same 3D CT scan and visually comparable to AP X-rays. Each image contains a red overlay representing the liver. The overlays differ between the two images.
Your task is to determine which overlay more accurately represents the liver.
Anatomical guidelines:
1. *Size and shape* — The liver is a large triangular or wedge-shaped organ; the overlay should reflect a broad smooth outline.
2. *Location* — It lies in the upper right abdomen (left side of the image in AP view), beneath the diaphragm and partially crossing the midline.
3. *Relation to rib cage* — Most of the liver is covered by ribs; the overlay should correspond to this region.

After evaluating both images, determine whether Image1 or Image2 contains the more anatomically accurate liver annotation.

Label Expert Prompt: Pancreas Annotation Comparison

You are given two images, Image1 and Image2, representing frontal projections of the same 3D CT scan and visually comparable to AP X-rays. Each image contains a red overlay representing the pancreas, and the overlays differ.
Your task is to determine which overlay more accurately represents the pancreas.
Anatomical guidelines:
1. *Shape* — The pancreas is an elongated organ with a thicker head and a thinner tail.
2. *Position* — It lies in the upper abdomen, posterior to the stomach and near the lower ribs, usually spanning horizontally with mild curvature.
3. *Continuity* — The pancreas should appear as one smooth, continuous structure without disconnected components.

After comparing both images, conclude whether Image1 or Image2 contains the more anatomically accurate pancreas annotation.

Label Expert Prompt: Rib Annotation Comparison (L1–L12 & R1–R12)

You are given two images, Image1 and Image2, representing frontal projections of the same 3D CT scan and visually comparable to AP X-rays. Each image contains red overlays marking the ribs (Left 1–12 and Right 1–12), which differ between the two images.
Your task is to determine which overlay more accurately represents the rib anatomy.
Anatomical guidelines:
1. *Completeness* — Twelve ribs should be present on each side, without missing or duplicated structures.
2. *Symmetry* — Corresponding rib pairs should have similar curvature and placement.
3. *Curvature* — Ribs should appear as smooth arcs extending laterally from the spine.
4. *Ordering* — The pattern should follow the natural superior-to-inferior rib order.
5. *Continuity* — Each rib should be a continuous curved structure.

After comparing both images, conclude whether Image1 or Image2 contains the more anatomically accurate rib annotation.

A.2. The Maximization Step

A.2.1. ROC Analysis for Tumor Annotation

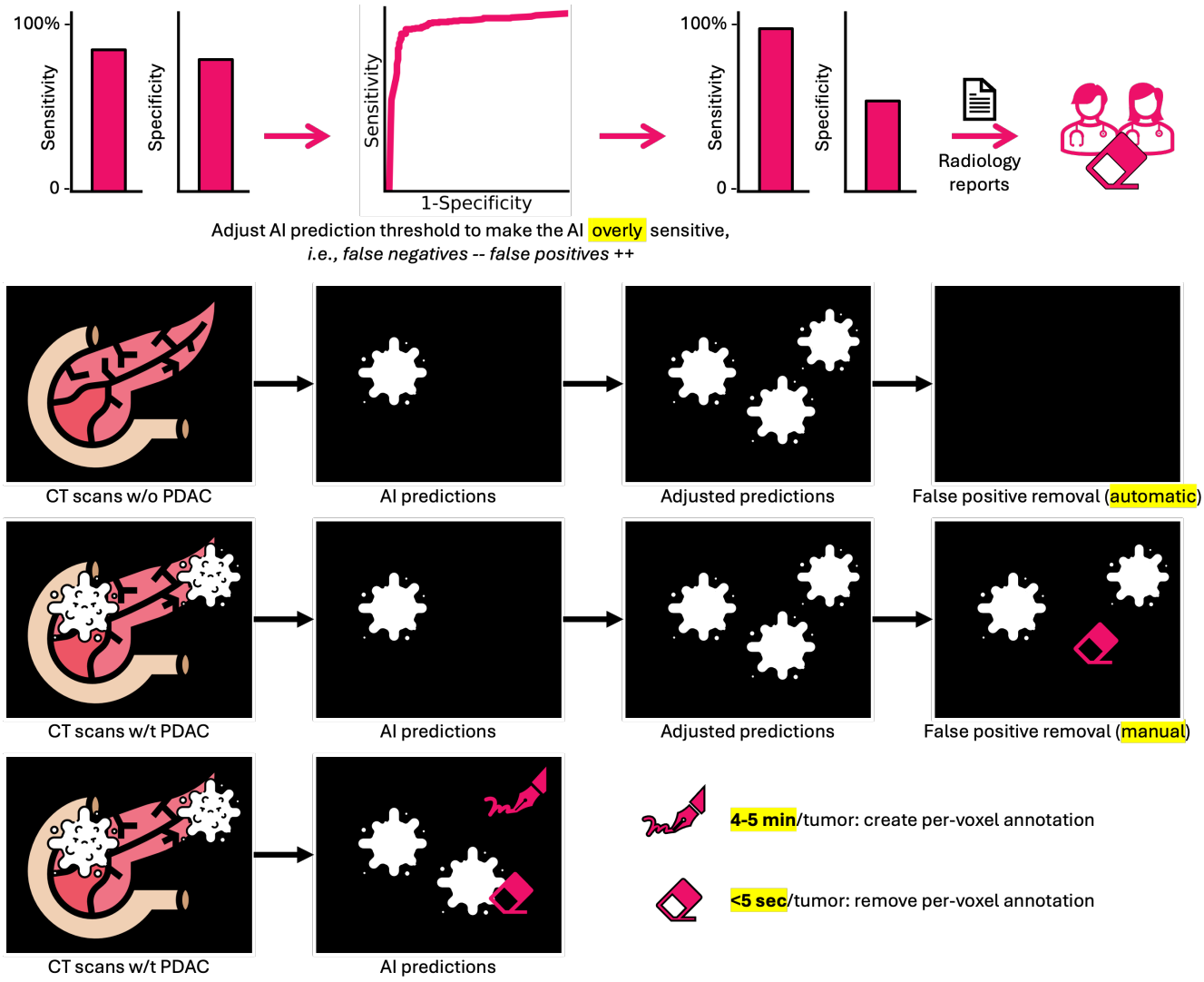

Figure 7. **ROC Analysis for Pancreatic Tumor Annotation.** We propose an efficient strategy, called ROC analysis, to assist radiologists in annotating tumors within a large-scale dataset (e.g., over 47,000 CT scans in our study). During the iterative data curation and annotation process facilitated by the ScaleMAI framework, AI model performance improves as data quality increases. In turn, stronger AI models generate more accurate pseudo labels with high sensitivity and specificity, significantly reducing radiologists' workload. Our observations show that removing AI false positives is much faster than creating per-voxel annotations for false negatives (missed tumors). Removing a false positive takes less than five seconds, whereas creating per-voxel annotations for a missed tumor can take 4–5 minutes. This insight motivates us to analyze the AI model's receiver operating characteristic (ROC) curve, which allows us to adjust the prediction threshold to prioritize sensitivity over specificity. To minimize radiologists' workload, we aim for nearly perfect sensitivity while maintaining acceptable specificity. By intentionally biasing the model towards high sensitivity, the AI minimizes missed tumors but inevitably introduces more false positives. Since handling false positives is simpler, this trade-off optimizes efficiency: (1) ***False positives in non-tumor CT scans*** can be automatically removed by cross-referencing radiology reports, which are typically available in clinical repositories (as illustrated in the second line in the Figure). (2) ***False positives in tumor CT scans*** can be efficiently removed using open-source annotation tools [11]. These tools enable radiologists to erase false positives with a few clicks, leveraging the AI's highly sensitive per-voxel predictions. In our study, we achieved 99% sensitivity for pancreatic tumor detection with only 0.6 false positives per scan. This means radiologists only have to remove just one false positive for every two CT scans (as illustrated in the third line in the Figure). Compared to creating per-voxel annotations from scratch, our ROC analysis approach reduces annotation time by up to 92%, significantly streamlining the workflow.

A.2.2. Continual Tuning

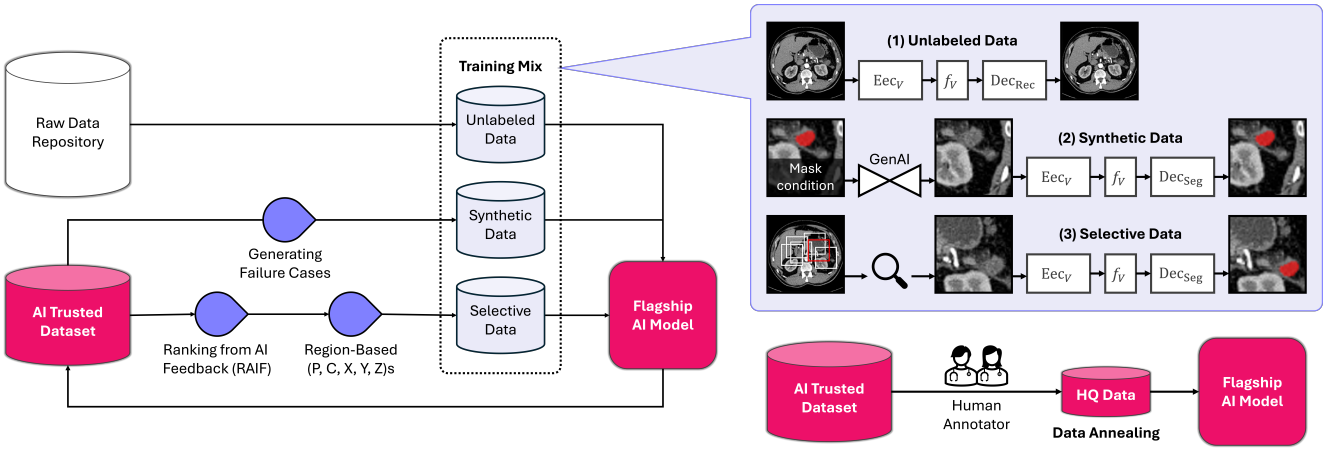

Figure 8. **Training Flagship Model with Data Mix and Data Annealing.** To optimize the training of Flagship Model, we incorporate a combination of data mix and data annealing strategies. The data mix consists of three primary types: ***First***, unlabeled data is utilized for self-supervised representation learning. This approach leverages the vast quantities of raw clinical data generated daily, requiring no manual annotation. The learned representations effectively regularize the model, enabling faster and more efficient learning of segmentation tasks with reduced reliance on annotated data. This methodology, supported by extensive literature, demonstrates the potential to exploit unlabeled clinical data for robust model training. ***Second***, synthetic data is employed to generate a diverse array of scans. These include variations across demographics, scanner types, and contrast enhancements, as well as tumors with differing locations, shapes, textures, sizes, and intensities that are not fully represented in the training set. This diversity enhances the model's robustness, particularly when encountering out-of-distribution test cases. ***Third***, selective data focuses on the most challenging regions of CT scans that confuse the model during training, as identified by the loss function. By prioritizing repeated sampling of these regions, the model learns more efficiently, avoiding the inefficiencies of processing non-informative areas such as air, bedding, or irrelevant anatomical regions. This targeted approach ensures that the model focuses on clinically relevant areas, such as the pancreas or abdominal region. ***Finally***, once the model is trained on data mix, we introduce data annealing to further fine-tune Flagship Model. We identify a gold-standard subset, consisting of voxel-level annotations meticulously created by expert radiologists. This data annealing technique has proven effective in large-scale training efforts in other domains, such as ChatGPT [1] and Llama 3 [20]. However, in the medical field, the lack of gold-standard data and the predominance of silver-standard annotation have limited its exploration. When releasing the dataset, we will explicitly mark this gold-standard subset to facilitate further research and development in the field.

B. Quality Assessment of PanTS-XL Datasets

B.1. Annotation Standard for 88 Anatomical Structures

The annotated areas of all tubular structures include both the tube wall and the lumen but exclude surrounding tissues, such as organs, mesentery, and adipose tissue. The pancreatic duct is identified as a low-attenuation tubular structure within the pancreas and should be marked from the tail to the ampulla of Vater. The common bile duct (CBD) appears as a low-attenuation tubular structure and should be annotated from the confluence of the common hepatic duct and bile duct to the ampulla of Vater. The superior mesenteric artery (SMA) is highlighted as a bright arterial structure originating from the aorta and should be traced from its origin to the point where it branches. The celiac artery is a short vessel arising from the aorta and splitting into the left gastric, splenic, and common hepatic arteries; it should be annotated from its origin to the bifurcation. Veins include the portal vein, splenic vein, and superior mesenteric vein: the portal vein is traced from the confluence of the splenic and superior mesenteric veins to its intrahepatic entry; the splenic vein is marked from the splenic hilum to the SMV confluence; and the superior mesenteric vein (SMV) is annotated from its major tributaries to the confluence with the splenic vein. Solid organs—including the liver, liver segments 1–8, pancreas (and its head, body, and tail), kidneys (left and right), adrenal glands (left and right), spleen, stomach, duodenum, intestine, colon, gallbladder, esophagus, bladder, prostate, and lungs (left and right)—are annotated by including the entire parenchyma while excluding surrounding fat, adjacent organs, and extrinsic vasculature. The aorta and postcava are annotated as continuous tubular structures throughout their visible abdominal course, following the same lumen-plus-wall rule as other vessels. For osseous structures, each rib (left and right, 1–12) is annotated to include the full cortical and cancellous bone while excluding costal cartilage; likewise, each vertebra—from C1–C7, T1–T12, and L1–L5—is annotated to include the vertebral body, pedicles, transverse processes, and spinous process while excluding intervertebral discs and surrounding soft tissues. The femur left and femur right are annotated by including the entire visible femoral head, neck, and proximal shaft, capturing both cortical and cancellous bone while excluding surrounding muscle and soft tissue. These standards ensure consistent and anatomically faithful annotations across all 88 structures in the dataset.

B.2. Reader Study: Tumor Detection & Diagnosis

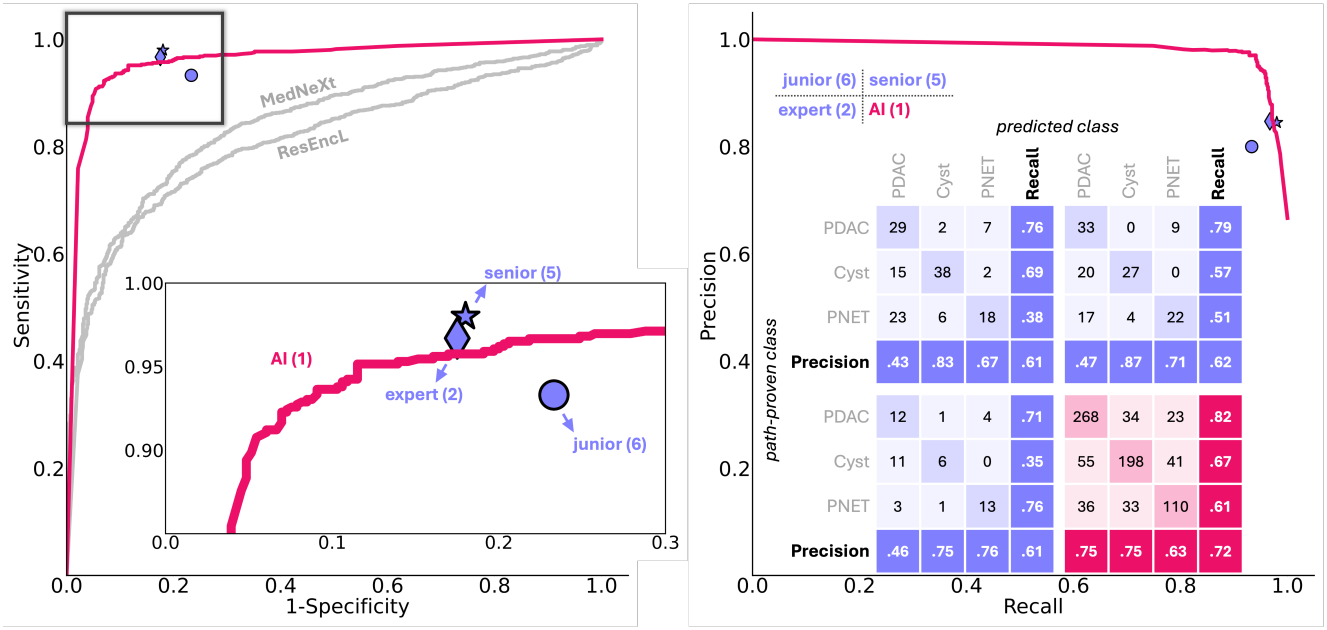

Figure 9. **Flagship Model matches senior and expert radiologists in tumor detection and surpasses them in tumor diagnosis accuracy.** We conducted an extensive multi-institution, multi-reader study comparing Flagship Model with radiologists with varying levels of experience. Thirteen board-certified radiologists were participated, including 6 juniors (<8 years of experience), 5 seniors (8–15 years), and 2 experts (>15 years). Each radiologist independently reviewed contrast-enhanced abdominal CT scans in the venous and arterial phases from 50 patients (100 CT scans), representing a broad spectrum of pancreatic conditions, including normal cases and tumors of three common subtypes: cysts, pancreatic adenocarcinoma (PDAC), and pancreatic neuroendocrine tumors (PNET). Radiologists were blinded to the proportion of normal and tumor cases and tasked with detecting and localizing tumors using 3D Slicer by marking any point within the tumor. They also classified tumors into the specified subtypes without access to patient medical history or symptom information. Flagship Model was evaluated under identical conditions on a larger cohort of 982 patients (1,964 CT scans). Performance was assessed using Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for tumor detection and confusion matrices for diagnosis. For tumor detection, Flagship Model (pink curve) achieved performance comparable to expert (blue diamond) and senior radiologists (blue star), outperforming junior radiologists (blue circle). In tumor diagnosis of PDAC, cysts, and PNET, Flagship Model achieved 72% accuracy, exceeding junior, senior, and expert radiologists by 11%, 10%, and 11%, respectively.

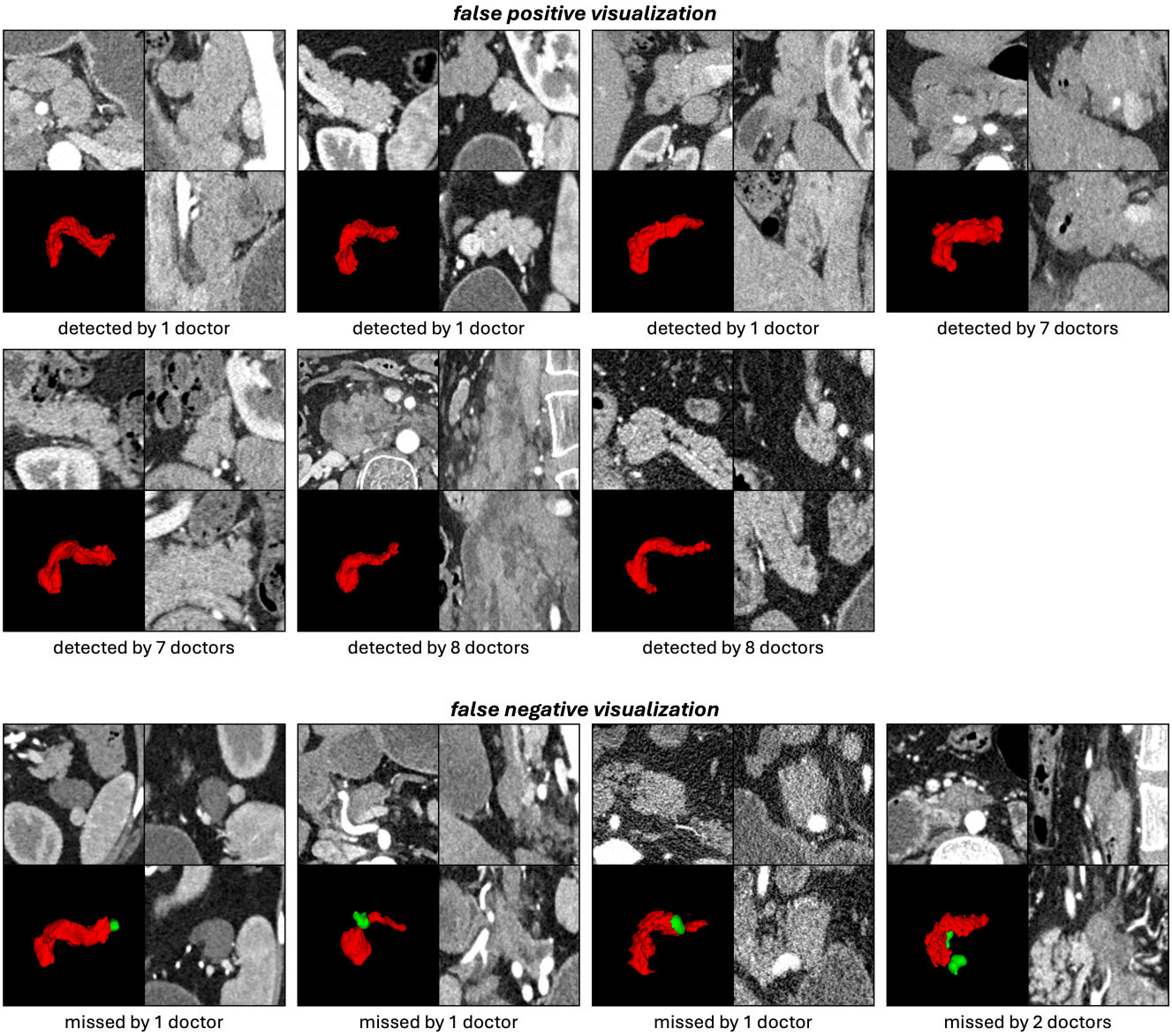

Figure 10. **Visualization of false positives and negatives predicted by radiologists.** In the false positive cases, the radiologists noticed slight irregularities in the pancreas tissue texture. However, these cases lacked two key reliable warning signs that typically indicate pancreatic tumor: abnormal widening of the main pancreatic duct and localized tissue shrinkage. The false negative cases demonstrated more subtle findings. One case showed a tumor growing outward from the tail end of the pancreas—a location that is often difficult for human readers if not examined thoroughly. In two other cases, while no obvious tumors were visible, there were areas where the pancreas tissue had become unusually thin, which often signals an underlying tumor in that location.

Table 4. **Flagship Model matches the pancreatic tumor detection performance of senior and expert radiologists, outperforming junior radiologists.** We present the sensitivity (%) and specificity (%) of pancreatic tumor detection across different tumor sizes—small (<20 mm), medium (20–40 mm), large (>40 mm), and all sizes—for radiologists at various career stages and Flagship Model. 13 radiologists of varying experience levels each evaluated 50 patients individually, while Flagship Model was tested on 982 patients. For all tumor sizes, all radiologist groups demonstrated high sensitivity, but only senior and expert radiologists achieved good specificity. Flagship Model surpassed all radiologist groups in specificity and had higher sensitivity than junior radiologists, approaching the performance of senior and expert radiologists. Notably, Flagship Model attained 100% sensitivity for medium and large tumors, suggesting it could assist all radiologists in detecting medium-sized tumors and help junior radiologists with large tumor detection. These findings indicate that Flagship Model performs at a level comparable to experienced radiologists, highlighting its potential as a reliable tool in clinical practice.

		Sensitivity, %				Specificity, %
career stage	reader	all-size	small (<20mm)	medium (20–40mm)	large (>40mm)	normal
junior (<8 years)	Reader 1	96.7 (29/30)	100 (10/10)	90.0 (9/10)	100 (10/10)	75.0 (15/20)
	Reader 2	100 (30/30)	100 (10/10)	100 (10/10)	100 (10/10)	75.0 (15/20)
	Reader 3	100 (30/30)	100 (10/10)	100 (10/10)	100 (10/10)	80.0 (16/20)
	Reader 4	96.7 (29/30)	100 (10/10)	90.0 (9/10)	100 (10/10)	80.0 (16/20)
	Reader 5	76.7 (23/30)	90.0 (9/10)	80.0 (8/10)	60.0 (6/10)	85.0 (17/20)
	Reader 6	90.0 (27/30)	100 (10/10)	90.0 (9/10)	80.0 (8/10)	65.0 (13/20)
	average	93.3 (168/180)	98.3 (59/60)	91.7 (55/60)	90.0 (54/60)	76.7 (92/120)
senior (8–15 years)	Reader 7	96.7 (29/30)	100 (10/10)	90.0 (9/10)	100 (10/10)	80.0 (16/20)
	Reader 8	100 (30/30)	100 (10/10)	100 (10/10)	100 (10/10)	80.0 (16/20)
	Reader 9	96.7 (29/30)	100 (10/10)	90.0 (9/10)	100 (10/10)	80.0 (16/20)
	Reader 10	100 (30/30)	100 (10/10)	100 (10/10)	100 (10/10)	80.0 (16/20)
	Reader 11	96.7 (29/30)	100 (10/10)	90.0 (9/10)	100 (10/10)	90.0 (18/20)
	average	98.0 (147/150)	100 (50/50)	94.0 (47/50)	100 (50/50)	82.0 (82/100)
expert (>15 years)	Reader 12	93.3 (28/30)	100 (10/10)	80.0 (8/10)	100 (10/10)	80.0 (16/20)
	Reader 13	100 (30/30)	100 (10/10)	100 (10/10)	100 (10/10)	85.0 (17/20)
	average	96.7 (58/60)	100 (20/20)	90.0 (18/20)	100 (20/20)	82.5 (33/40)
	Flagship Model	94.1 (640/680)	85.2 (468/549)	100 (105/105)	100 (10/10)	83.8 (253/302)

B.3. PanTS-XL vs. Public Organ and Tumor Datasets

Table 5. **Comparison of public organ and pancreatic tumor CT datasets.** PanTS-XL surpasses prior organ and pancreatic CT datasets in both scale and completeness, spanning 47,315 CTs across 112 centers and 13 countries. It integrates heterogeneous scanners, full demographic metadata (age, sex, phase), and clinically meaningful labels (narrative and structured reports), forming the most comprehensive foundation for scalable medical intelligence to date.

	Scale						Metadata			Clinical	
Dataset	# of CT	# of new CT	# of structure	# of annotation	# of center	# of country	age	sex	phase	narrative report	structured report
CHAOS [2018]	40	40	1	40	1	1	✗	✗	✗	✗	✗
BTCV [2015]	50	50	12	600	1	1	✗	✗	✗	✗	✗
CT-ORG [2020]	140	140	6	840	8	6	✗	✗	✗	✗	✗
WORD [2021]	150	150	16	2400	1	1	✗	✗	✗	✗	✗
LiTS [2019]	201	201	1	201	7	5	✗	✗	✗	✗	✗
AMOS22 [2022]	500	500	15	7500	2	1	✗	✗	✗	✗	✗
KiTS [2019]	599	599	1	599	1	1	✗	✗	✗	✗	✗
AbdomenCT-1K [2021]	1,112	✗	4	4448	12	7	✗	✗	✗	✗	✗
TotalSegmentator [2023]	1,228	1,228	117	143,676	10	1	✔	✔	✔	✗	✗
FLARE'23 [2024]	4,100	✗	13	53,300	30	✗	✗	✗	✗	✗	✗
Trauma Detect. [2024]	4,711	4,711	✗	✗	23	12	✗	✗	✗	✗	✗
AbdomenAtlas [2025, 2025, 2024]	9,262	✗	25	231,550	89	17	✔	✔	✔	9,262	9,262
TCIA–panNET [2023]	38	38	1	38	1	1	✔	✔	✔	✗	✗
TCIA-Pancreas [2021]	82	82	1	82	1	1	✗	✗	✔	✗	✗
MSD–Pancreas [2022]	420	420	1	420	1	1	✗	✗	✔	✗	✗
PANORAMA [2024]	2,238	1,964	5	11,190	5	3	✔	✔	✗	✗	✗
PanTS [2025]	9,901	556	27	267,327	76	12	✔	✔	✔	✗	9,901
PanTS-XL (Ours)	**47,315**	**12,000**	**88**	**4,163,720**	**112**	**13**	✔	✔	✔	**47,315**	**47,315**

C. Experimental Results of Flagship Model

C.1. Benchmarking Flagship Model: Evaluation Metrics

Diagnosis. For the tumor detection task in the binary classification setting, we use Sensitivity, Specificity and F1-score as our evaluation metrics. Let TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. We define sensitivity (also called recall) as the fraction of actual positives that are correctly identified:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{1}$$

Specificity measures the fraction of actual negatives that are correctly identified:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{2}$$

The F1-score[4] combines precision (Precision $= \frac{\text{TP}}{\text{TP+FP}}$) and recall (Recall $= \frac{\text{TP}}{\text{TP+FN}}$) into a single metric, and is defined as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

Segmentation. For the tumor segmentation task, we use the Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD) as our evaluation metrics. The DSC compares the overlap between two sets, commonly the predicted segmentation A and the ground-truth segmentation B. It is given by:

$$\text{DSC} = \frac{2|A \cap B|}{|A| + |B|}. \tag{4}$$

The NSD metric quantifies how closely the surfaces of the predicted and ground-truth segments match within a specified tolerance δ. Let S_p and S_g denote the sets of boundary points for the predicted and ground-truth surfaces, respectively, and let $d(x, S)$ represent the minimum distance from a point x to any point in set S. The NSD is then defined as the proportion of boundary points (from both S_p and S_g) that lie within δ of each other:

$$\text{NSD} = \frac{\sum_{x \in S_p} \mathbf{1}[\, d(x, S_g) < \delta] \; + \; \sum_{x \in S_g} \mathbf{1}[\, d(x, S_p) < \delta]}{|S_p| + |S_g|}, \tag{5}$$

where $\mathbf{1}[\cdot]$ is the indicator function, equal to 1 if the condition is satisfied and 0 otherwise.

[4]The F1-score combines sensitivity (recall) and precision, providing a balanced metric for scenarios involving class imbalance.

C.2. Benchmarking Flagship Model: Dataset Attributes

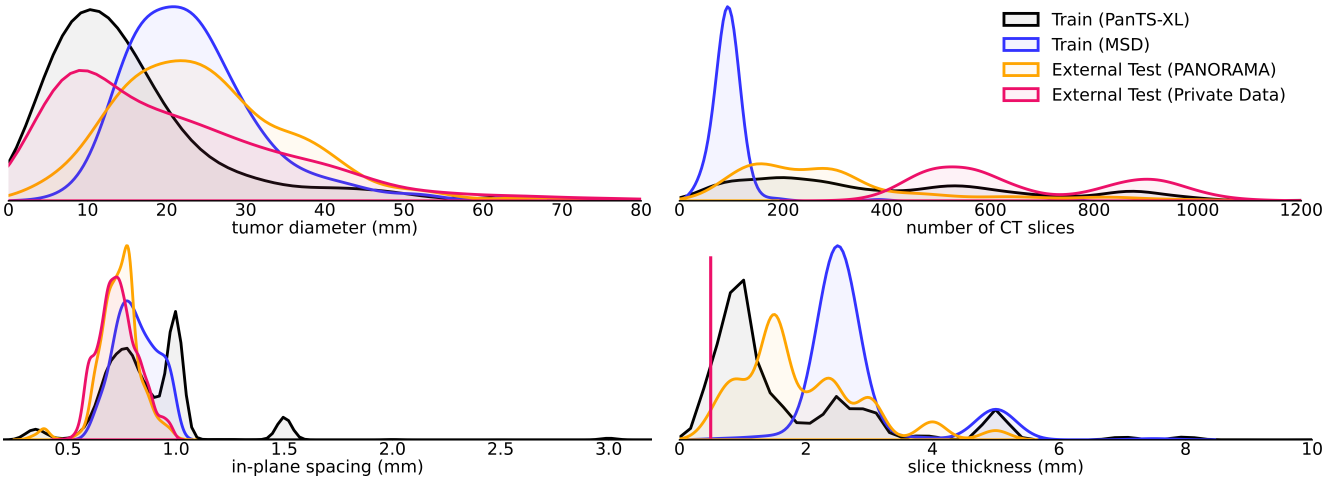

Figure 11. **Dataset attributes.** Our study used four datasets: two for training AI models and two for testing them. Before the creation of PanTS-XL (47,315 CT scans), the publicly available MSD-Pancreas dataset was the *only* resource for training pancreatic tumor segmentation models. Therefore, all baseline AI models in this study were trained on MSD-Pancreas. Existing literature suggests that AI is vulnerable when applied to CT scans from datasets with differing attributes [23, 65], such as variations in tumor diameter, the number of CT slices, in-plane spacing, and slice thickness. To evaluate the robustness of the baseline models, we conducted external validation using two datasets sourced from hospitals distinct from those contributing to MSD-Pancreas (Memorial Sloan Kettering Cancer Center, USA). The first external dataset, PANORAMA, was sourced from five hospitals across three countries, including Dutch, Sweden and Norway [3]. The second dataset, a proprietary dataset, was gathered from hospitals in a country distinct from MSD-Pancreas. As seen, compared with MSD-Pancreas, the PANORAMA dataset differs in the number of CT slices and slice thickness, while the proprietary dataset diverges significantly across all four attributes. The test results in Table 2 reveal that baseline models perform better on the PANORAMA, with the proprietary dataset yielding a much lower performance. These findings validate the hypothesis that discrepancies between training and test data significantly impact AI performance and robustness. This motivated us to create PanTS-XL that offers a substantially larger training set (47,000 annotated CT scans vs. MSD-Pancreas's 200) with greater diversity in key attributes as illustrated by the black curve.

C.3. PDAC, Cyst, and PNET Diagnosis (+7% Accuracy)

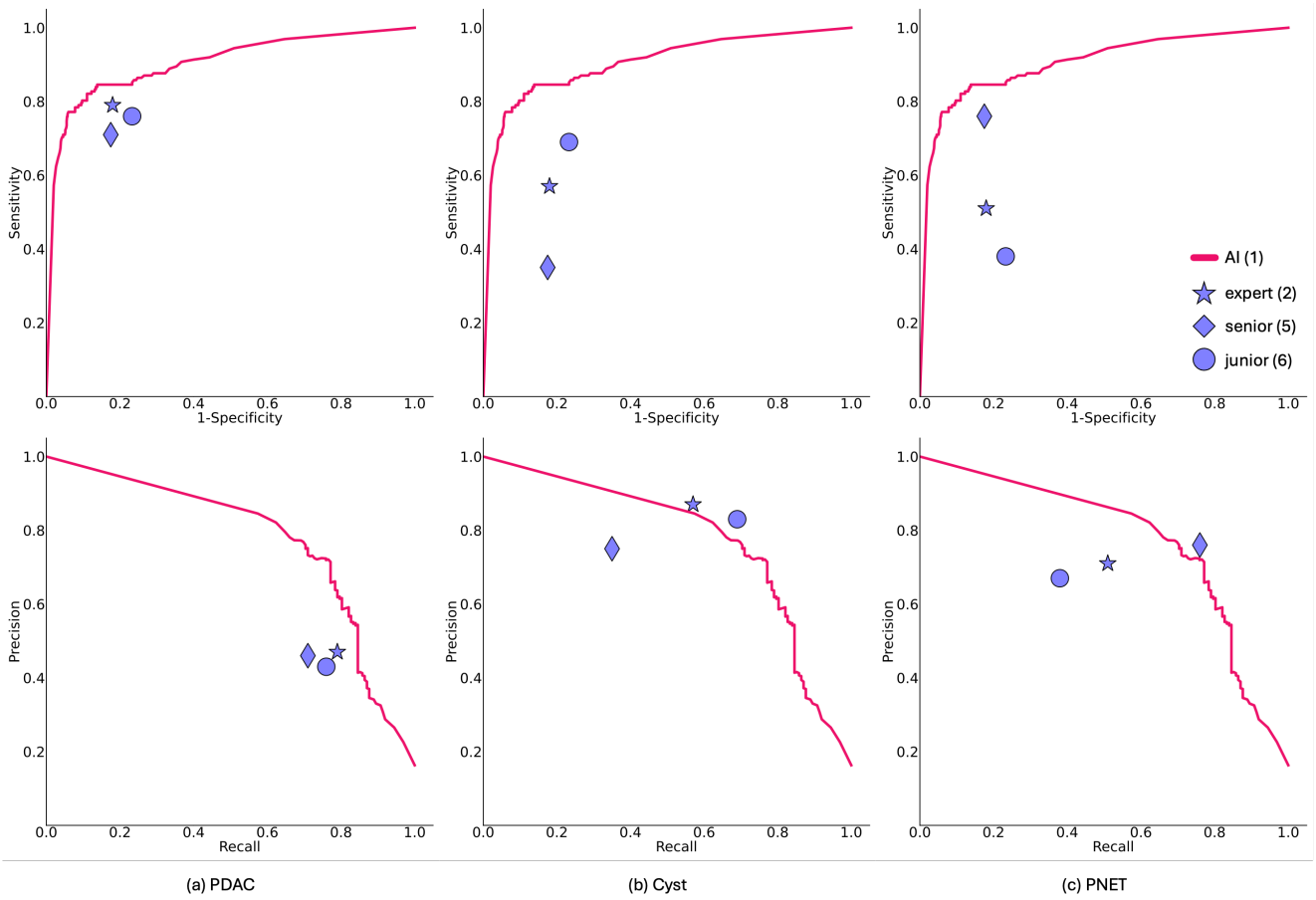

Figure 12. **Our Flagship Model can approach radiologists' performance in pancreatic tumor diagnosis.** Classifying pancreatic tumor types (Cyst, PDAC, PNET) directly from CT scans is challenging due to the subtle and overlapping visual features among these tumors. Key characteristics such as shape, size, and enhancement patterns can vary significantly within the same tumor type and often mimic those of other types. Additionally, CT scans lack the biological and molecular context provided by patient symptoms, medical history, follow-up imaging, or biopsy results, which are crucial for accurate diagnosis. Furthermore, variations in imaging protocols and scanner settings across institutions add complexity, making it difficult for both radiologists and AI models to achieve high accuracy. Using our annotated dataset of tumor types, this study marks the first time that: (1) AI performance is evaluated on a publicly available dataset, enabling reproducibility. (2) Radiologists are tested on the same dataset, allowing others to benchmark their performance. (3) AI is directly compared with radiologists across different career stages.

C.4. Pancreatic Tumor Detection (+10% Sensitivity)

Table 6. **Flagship Model, with a backbone of ResEncL, achieves the best performance for pancreatic tumor detection.** Note that these are tumor-wise detection results. Performance is given as sensitivity, specificity, and F1-score. Best-performing results are **bolded** for each dataset. In addition, we have performed a one-sided Wilcoxon signed rank test between the best-performing model and others [71]. The performance gain is statistically significant at the $P = 0.05$ level, with highlighting in a pink box.

		PANORAMA (N=1,964)†	Proprietary dataset (N=1,958)		
method	training set	Sensitivity	Sensitivity	Specificity	F1-score
Swin UNETR [66]	MSD-Pancreas	85.5 (497/581)	25.2 (824/3426)	16.7 (104/623)	35.9 (1728/4809)
UniSeg [73]	MSD-Pancreas	77.8 (452/581)	23.4 (801/3426)	78.5 (485/623)	36.7 (1602/4361)
ResEncL [33]	MSD-Pancreas	74.7 (434/581)	23.7 (813/3426)	87.0 (542/623)	38.1 (1660/4360)
STU-Net-Base [31]	MSD-Pancreas	76.8 (446/581)	23.1 (791/3426)	84.4 (526/623)	36.7 (1582/4314)
MedNeXt [61]	MSD-Pancreas	73.3 (426/581)	24.6 (842/3426)	85.2 (531/623)	39.3 (1726/4394)
DTI [46]	PANORAMA	–	26.7 (916/3426)	88.1 (549/623)	41.4 (1832/4416)
Flagship Model	PanTS-XL	**89.2** (518/581)	**40.2** (1381/3426)	**88.3** (550/623)	**56.6** (2754/4865)
Δ		**+3.7**	**+13.5**	**+0.2**	**+17.3**

†PANORAMA annotates only PDAC, treating all other types of pancreatic tumors and healthy pancreases as *Normal*—specificity and F1-score cannot be computed.

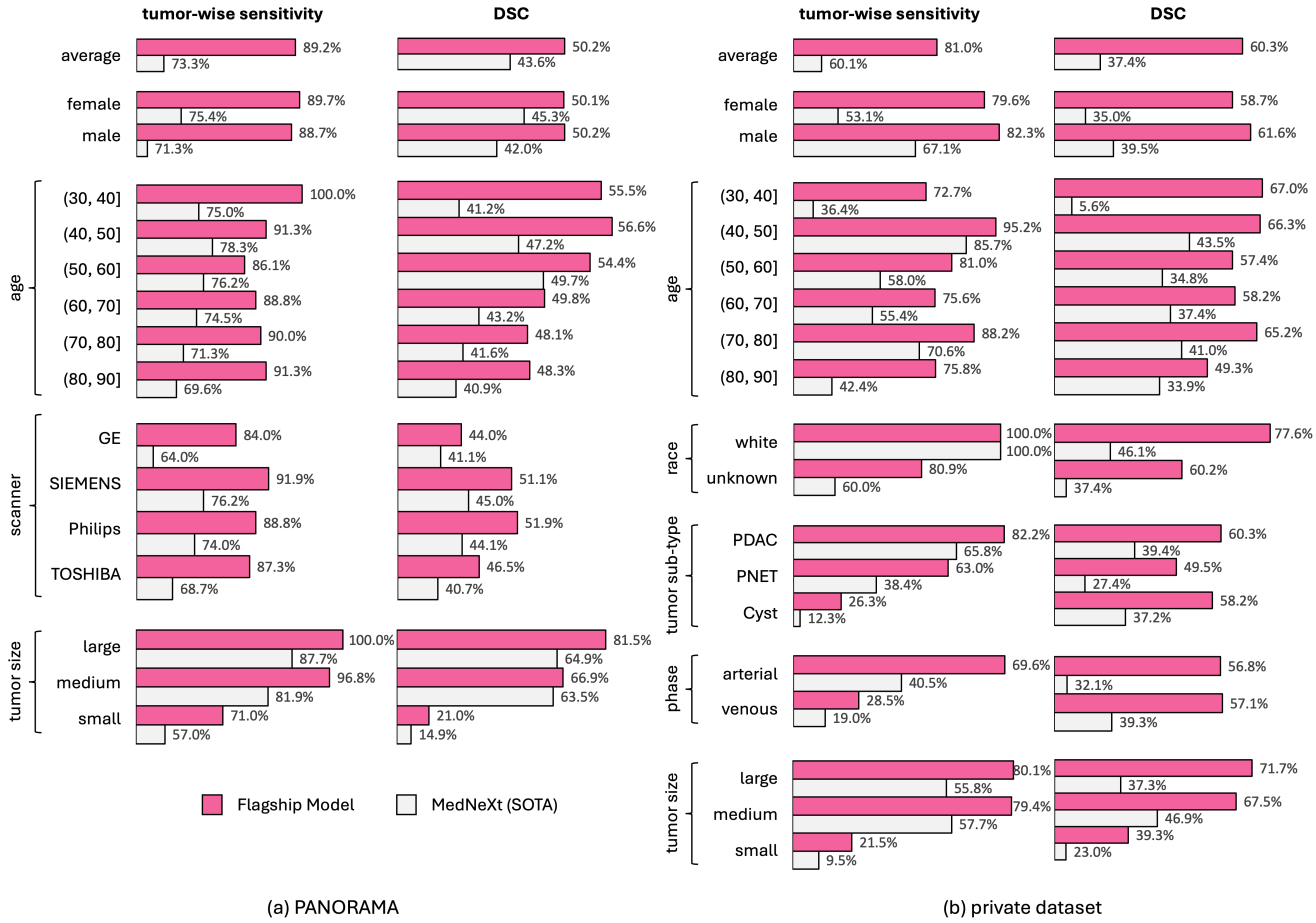

Figure 13. **Flagship Model demonstrates robust generalizability across diverse demographic and technical variations in out-of-distribution evaluations.** Flagship Model shows improved tumor detection and segmentation performance across various age, sex, race, tumor sub-types, tumor sizes, scanner types, and imaging phases. Flagship Model, trained on PanTS-XL (large-scale but silver standard), consistently matches or surpasses the performance of the MedNeXt model (top 1), which was trained on a smaller but gold standard dataset.

C.5. Pancreatic Tumor Segmentation (+14% DSC)

Table 7. **Flagship Model demonstrates robust generalizability across various demographic groups and scanner types in tumor segmentation.** We compare the median DSC and interquartile range (IQR) of Flagship Model and public top-performing MedNeXt model on the PANORAMA and proprietary datasets for sex, age, scanner type, and race. Notably, Flagship Model consistently achieves higher median DSC with statistically significant improvements (p-value$<$0.001 in most cases). For example, in the age group 70–80, Flagship Model achieved a median DSC of 73.9% compared to MedNeXt's 47.8%, a difference of 16.1% (p-value$<$0.001).

	Median DSC (IQR) on PANORAMA, %				Median DSC (IQR) on proprietary dataset, %			
group	Flagship Model	MedNeXt (SOTA)	difference	p-value	Flagship Model	MedNeXt (SOTA)	difference	p-value
all test samples	58.1 (24.4–76.2)	54.4 (0.0–74.8)	0.1 (-11.2–20.1)	$<$0.001	70.4 (44.4–81.3)	38.8 (0.0–67.7)	15.2 (2.7–40.8)	$<$0.001
sex								
female	57.5 (25.0–77.0)	55.9 (0.6–76.3)	0.0 (-12.7–18.7)	0.072	68.8 (39.8–79.8)	31.4 (0.0–68.0)	14.0 (1.5–41.2)	$<$0.001
male	58.9 (23.9–74.8)	53.4 (0.0–72.9)	0.7 (-9.8–21.0)	0.002	71.4 (49.2–82.5)	45.7 (0.3–67.5)	15.5 (3.5–40.1)	$<$0.001
age								
30–40	66.6 (40.4–81.7)	44.0 (12.9–72.3)	4.9 (-1.4–20.6)	0.608	70.5 (63.6–73.2)	0.1 (0.0–12.5)	62.8 (54.6–69.9)	$<$0.001
40–50	66.9 (52.9–82.3)	57.4 (8.8–74.2)	0.6 (-2.8–15.8)	0.325	74.8 (65.0–84.0)	54.0 (16.3–68.9)	17.7 (7.6–58.2)	0.012
50–60	63.5 (31.1–80.2)	60.0 (16.9–77.9)	0.0 (-14.6–19.2)	0.299	65.4 (42.1–78.6)	36.5 (0.0–62.7)	16.0 (1.1–42.4)	$<$0.001
60–70	57.5 (26.2–75.9)	54.1 (0.0–71.9)	0.0 (-11.7–20.8)	0.038	68.0 (38.0–81.5)	41.1 (0.0–66.8)	13.8 (3.1–35.7)	$<$0.001
70–80	55.8 (21.6–74.5)	50.1 (0.0–74.4)	0.0 (-11.1–19.2)	0.039	73.9 (57.0–83.1)	47.8 (0.7–71.9)	16.1 (3.8–40.6)	$<$0.001
80–90	55.4 (24.2–70.7)	44.2 (0.0–75.4)	0.4 (-6.3–25.0)	0.263	57.3 (9.3–78.7)	31.7 (0.0–66.8)	8.2 (0.0–18.7)	$<$0.001
scanner								
GE	47.1 (20.4–69.4)	54.5 (0.0–74.2)	0.0 (-4.1–10.0)	0.754	-	-	-	-
SIEMENS	57.7 (26.4–76.9)	57.3 (0.8–75.7)	0.5 (-12.9–21.9)	0.072	-	-	-	-
Philips	59.8 (30.0–75.9)	55.4 (0.0–72.0)	0.2 (-11.6–22.4)	0.006	-	-	-	-
TOSHIBA	57.3 (16.1–75.4)	44.1 (0.0–75.5)	0.0 (-8.3–17.6)	0.157	-	-	-	-
race								
white	-	-	-	-	77.6 (76.6–78.5)	46.1 (43.0–49.3)	31.4 (29.3–33.6)	0.102
unknown	-	-	-	-	70.3 (44.0–81.3)	38.6 (0.0–67.8)	15.0 (2.5–40.9)	$<$0.001