Segment anything in medical images (MedSAM)

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, Bo Wang (Toronto, ON, Canada)

Training script, inference script, and trained model: **https://github.com/bowang-lab/MedSAM**
https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-024-44824-z/MediaObjects/41467_2024_44824_MOESM1_ESM.pdf

Context: Segmentation fondation models

- **SAM = segment anything model,** remarkable versatility and performance across various segmentation tasks

Advantages:

- Wide range of segmentation tasks for one training
- Could lead to more consistent results across different tasks

But in medical imaging…

Challenges:

- Significant differences between image modalities (3D/2D), imaging protocols and machines, etc.

- Segmentation task can vary depending on the specific clinical scenario

  (Example: segmenting the liver tumor VS entire liver and surroundings organs on CT scan)

  - Need of user provided prompts

MedSAM training and validation method

Fine-tuning SAM with:

- **1,570,263 image-mask pairs** (unprecedented dataset; 80% training, 10% tuning, and 10% validation)
  - MRI/CT slices in **NifTI format** + [0, 255] intensity normalization + [1024×1024×3] resizing
  - Grayscale/RGB in **png format** + [1024×1024×3] resizing
- **10 images modalities**, with a multitude of imaging protocols
- **30 cancer types**
- **86 internal validation tasks**
- **60 external validation tasks**
  - Unseen segmentation targets
  - New domains (cell segmentation in LM images and organelle segmentation in EM images)

Comparison with:

- SOTA segmentation foundation model **SAM**
- 10 dedicated specialist models based on **U-Net** and **DeepLabV3+** respectively

MedSAM datasets distribution and MedSAM architecure

a

b

Input Image

Image encoder

Image embedding

Mask decoder
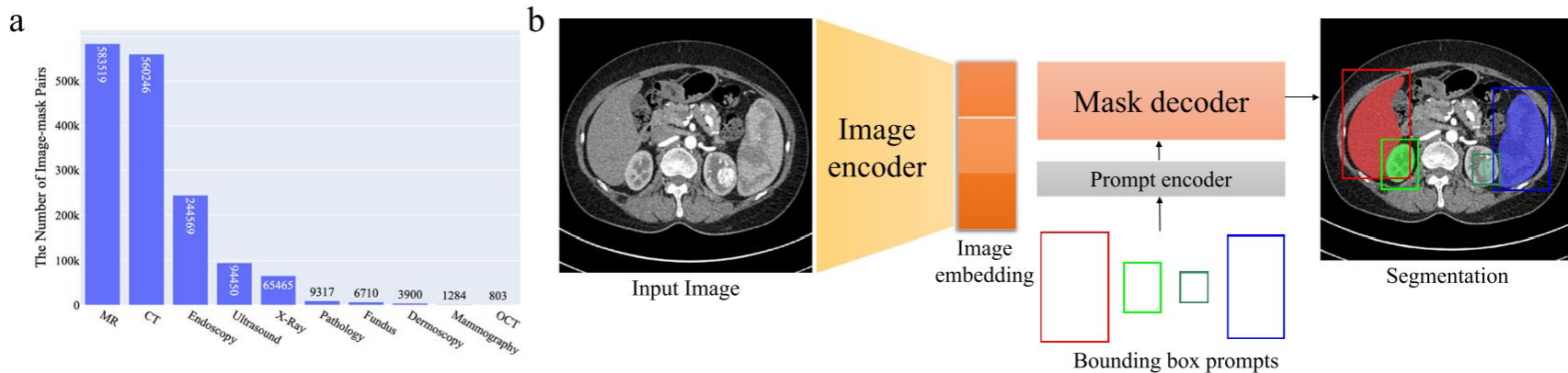
Prompt encoder

Bounding box prompts

Segmentation

Fig. 2 | Overview of the modality distribution in the dataset and the network architecture. a The number of medical image-mask pairs in each modality. **b** MedSAM is a promptable segmentation method where users can use bounding boxes to specify the segmentation targets. Source data are provided as a Source Data file.
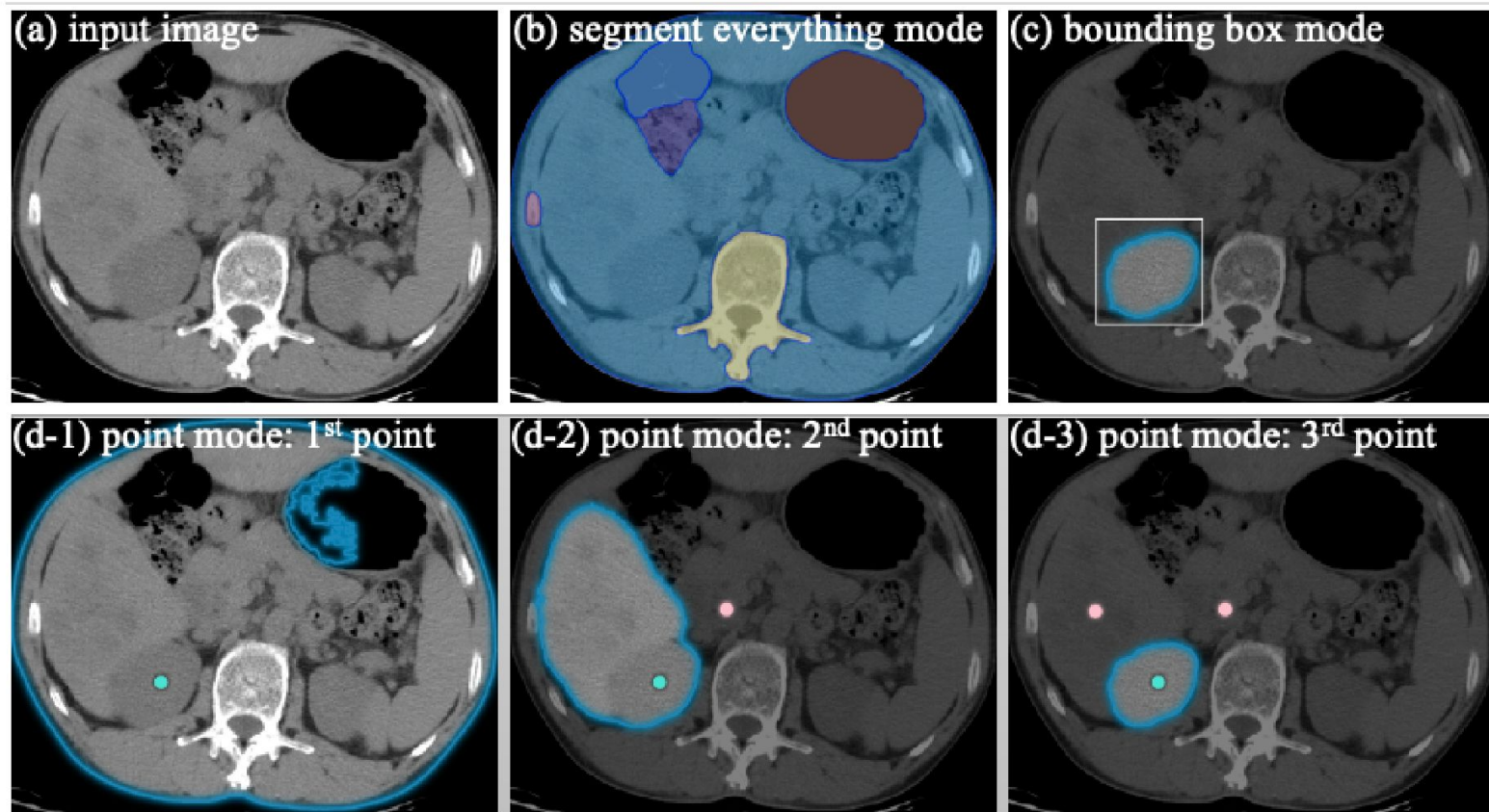
TABLE 5
The number of training images for modality-wise specialist models.

Modality	Num. of Training Images
Computed Tomography	362,229
Magnetic Resonance Imaging	442,818
X-Ray	36915
Mammography	986
Optical Coherence Tomography	642
Ultrasound	74761
Dermoscopy	2955
Endoscopy	194805
Fundus	960
Pathology	6239

- Image encoder based on vision transformer:
  - 12 transformer layers
  - multi-head self-attention and multilayer perceptron per block
- Mask decoder:
  - feature size in image embedding of 64×64

MedSAM prompt types

(a) input image

(b) segment everything mode

(c) bounding box mode

(d-1) point mode: 1st point

(d-2) point mode: 2nd point

(d-3) point mode: 3rd point

4

MedSAM Loss function

Binary cross entropy loss:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[g_i \log s_i + (1 - g_i)\log(1 - s_i)\right]$$

Dice loss:

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} g_i s_i}{\sum_{i=1}^{N}(g_i)^2 + \sum_{i=1}^{N}(s_i)^2}$$

Total loss:

$$L = L_{\text{BCE}} + L_{\text{Dice}}$$

si, gi denotes the predicted segmentation and ground truth of voxel i, respectively.
N is the number of voxels in the image I.

MR image datasets

TABLE 2

Magnetic Resonance (MR) image datasets. Datasets marked with ∗ denote external validation sets and the remaining datasets are used for internal validation.

Dataset Name	Modality	Segmentation Targets	# of scans	
ACDC* [30]	MR	Heart anatomies	150	https://humanheart-project.creatis.insa-lyon.fr/
AMOS-MR [6]	MR	Abdominal organ	40	https://amos22.grand-challenge.org/Dataset/
ATLAS R2.0 [31]	MR-T1	Brain stroke	1271	https://atlas.grand-challenge.org/
Brain Tumor Dataset Figshare [32], [33]	MR-T1ce	Brain tumor	233	https://www.kaggle.com/datasets/ashkhagan/
Brain TR-GammakKnife [34]	MR	Brain lesion	47	https://doi.org/10.7937/xb6d-py67
BraTS [35]–[39]	MR-T1, MR-T1CE, MR-T2, MR-FLAIR	Brain tumor	1251	http://braintumorsegmentation.org/
CC-Tumor Heterogeneity* [40]	MR	Cervical cancer	7	https://doi.org/10.7937/ERZ5-QZ59
CHAOS* [41]	MR-T1, MR-T2	Liver, kidney, spleen	60	https://chaos.grand-challenge.org/
crossMoDA [42]	MR	Brain tumor	227	https://crossmoda-challenge.ml/
FeTA [42]	MR-Fetal	Brain tissues	160	https://feta.grand-challenge.org/
HaN-Seg* [43]	MR	Head organs	42	https://zenodo.org/record/
ISLES [44]	MR-DWI, MR-ADC, MR-FLAIR	Ischemic stroke lesion	180	http://www.isles-challenge.org/
I2CVB [45]	MR-T2, MR-DWI	Prostate	19	https://i2cvb.github.io/
Meningioma-SEG-CLASS [46]	MR-T1ce, T2-FLAIR	Tumor (meningioma)	191	https://doi.org/10.7937/0TKV-1A36
MMs [47]	MR	Heart anatomies	150	https://www.ub.edu/mnms-2/
MSD-Heart [48]	MR	Left atrial	30	http://medicaldecathlon.com/
MSD-Prostate [21]	MR-ADC, MR-T2	Prostate	48	http://medicaldecathlon.com/#tasks
NCI-ISBI [49]	MR-ADC, MR-T2	Prostate	48	http://dx.doi.org/10.7937/K9/TCIA.2015.zF0vl
PI-CAI [50]	MR-bp	Prostate cancer	1584	http://github.com/DIAGNijmegen/picai_labels
PPMI [51]	MR-T1	Brain regions of Parkinson patients	1130	https://www.ppmi-info.org/access-data-specim
PROMISE [52]	MR-T2	Prostate	50	https://promise12.grand-challenge.org/Details/
Qin-Prostate-Repeatability [21], [53]	MR	Left atrium	30	http://doi.org/10.7937/K9/TCIA.2018.MR1CK(
QUBIQ* [54]	MR	Prostate	52	https://qubiq21.grand-challenge.org/
Spine [55]	MR	Vertebrae	172	https://www.cg.informatik.uni-siegen.de/en/sp
WMH [56]	MR-T1, MR-FLAIR	White matter hyper-intensities	60	https://wmh.isi.uu.nl/

Evaluation metrics

$$\text{DSC}(G,\,S) = \frac{2|G \cap S|}{|G| + |S|}$$

Dice similarity coefficient: evaluate the region overlap between expert annotation masks and segmentation results

$$\text{NSD}(G,\,S) = \frac{|\partial G \cap B_{\partial S}^{(\tau)}| + |\partial S \cap B_{\partial G}^{(\tau)}|}{|\partial G| + |\partial S|}$$

**Normalized surface distance**: evaluate the boundary consensus between expert annotation masks and segmentation results at a given tolerance

where $B_{\partial G}^{(\tau)} = \{x \in R^3 \mid \exists \tilde{x} \in \partial G,\ ||x - \tilde{x}|| \leq \tau\}$,

$B_{\partial S}^{(\tau)} = \{x \in R^3 \mid \exists \tilde{x} \in \partial S,\ ||x - \tilde{x}|| \leq \tau\}$ denote the border region of the expert

annotation mask and the segmentation surface at tolerance τ, respectively. In this paper, we

set the tolerance τ as 2.

Performance distribution of 86 internal validation tasks (DSC score)
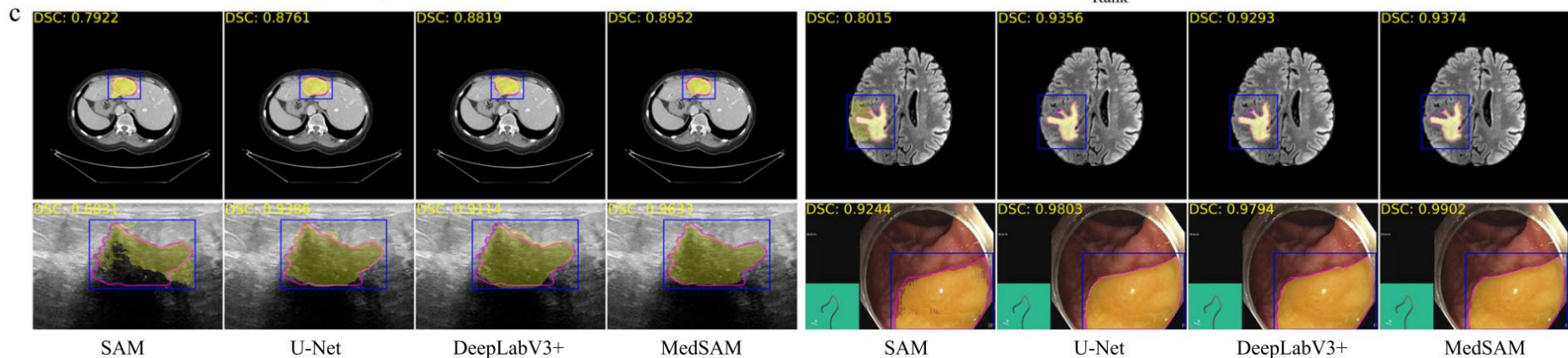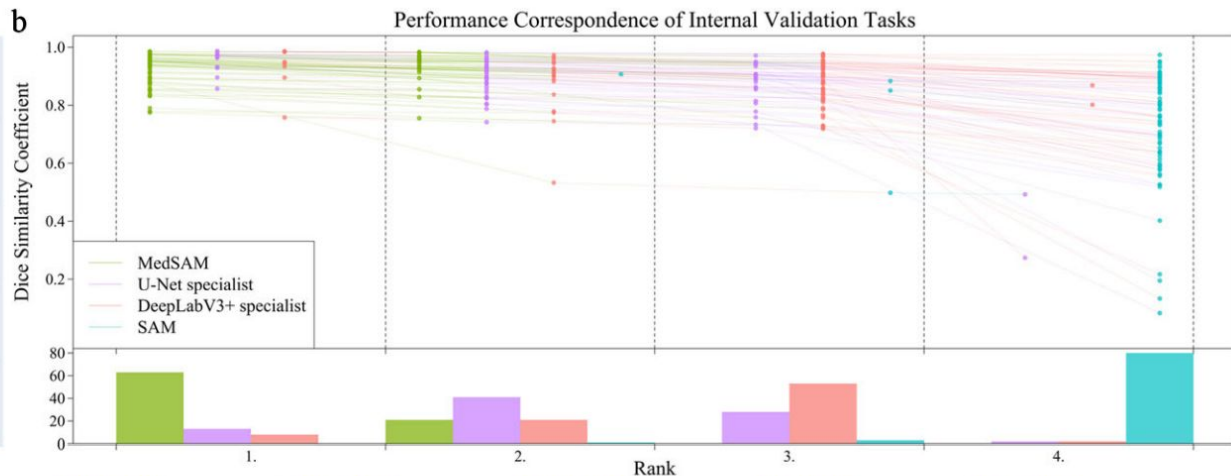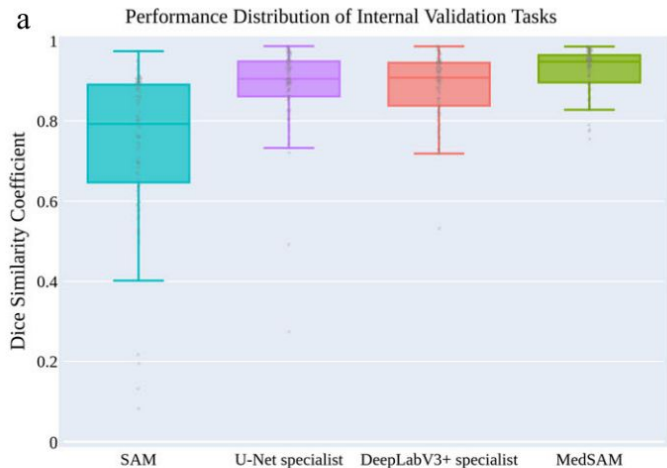
Performance distribution of 60 external validation tasks (DSC score)
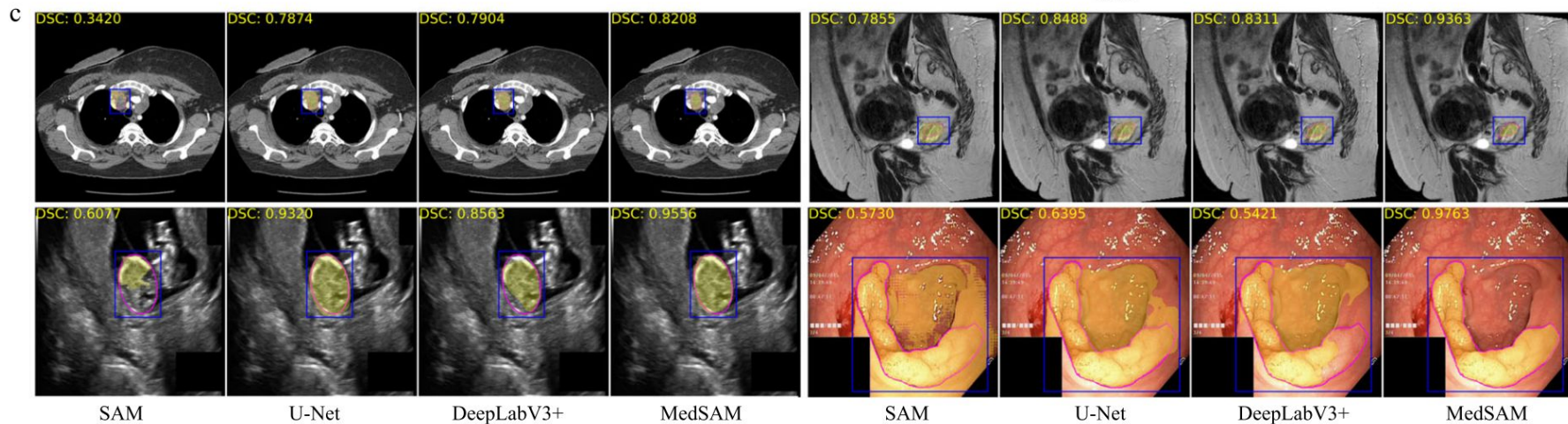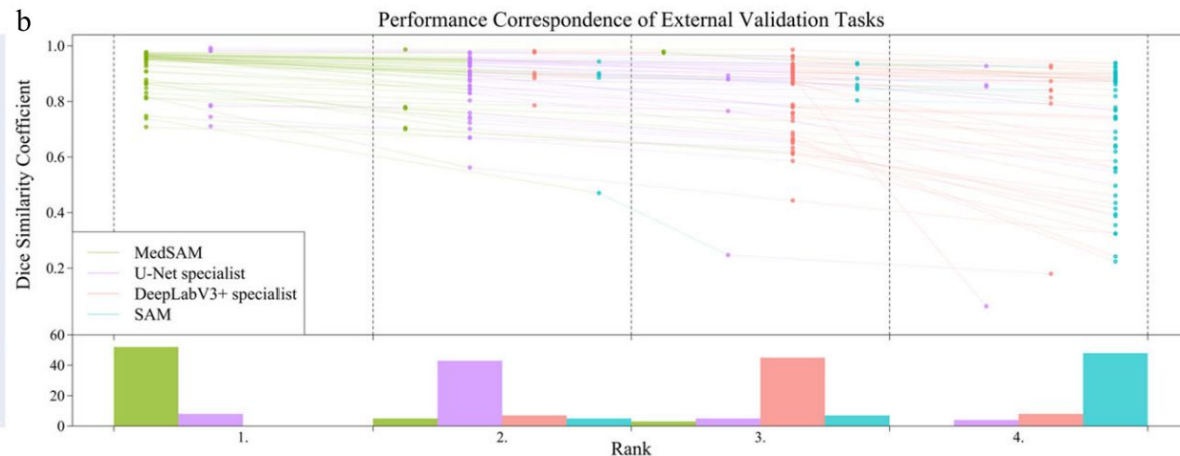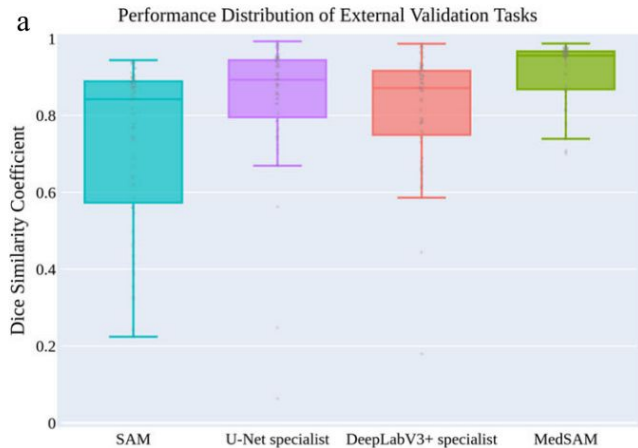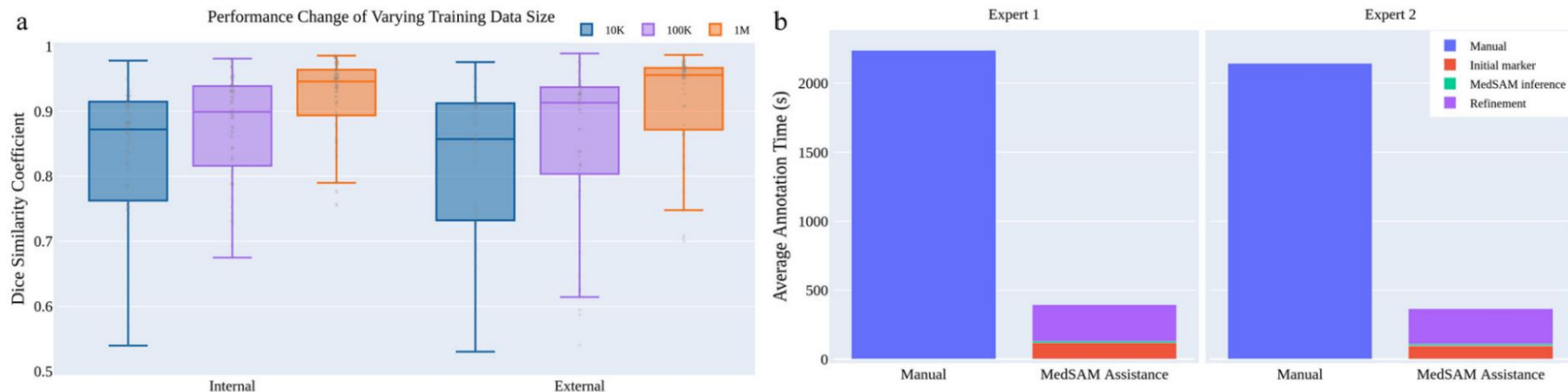
Context: Partial volume effect VS segmentation

Fig. 5 | The effect of training dataset size and a user study of tumor annotation efficiency. a Scaling up the training image size to one million can significantly improve the model performance on both internal and external validation sets.

b MedSAM can be used to substantially reduce the annotation time cost. Source data are provided as a Source Data file.

- Linear marker every 3-10 slices by two experts, before MedSAM segmentation, and expert revision (3D adrenal tumors)
- Annotation time reduced by 82.37% and 82.95% for the two experts respectively

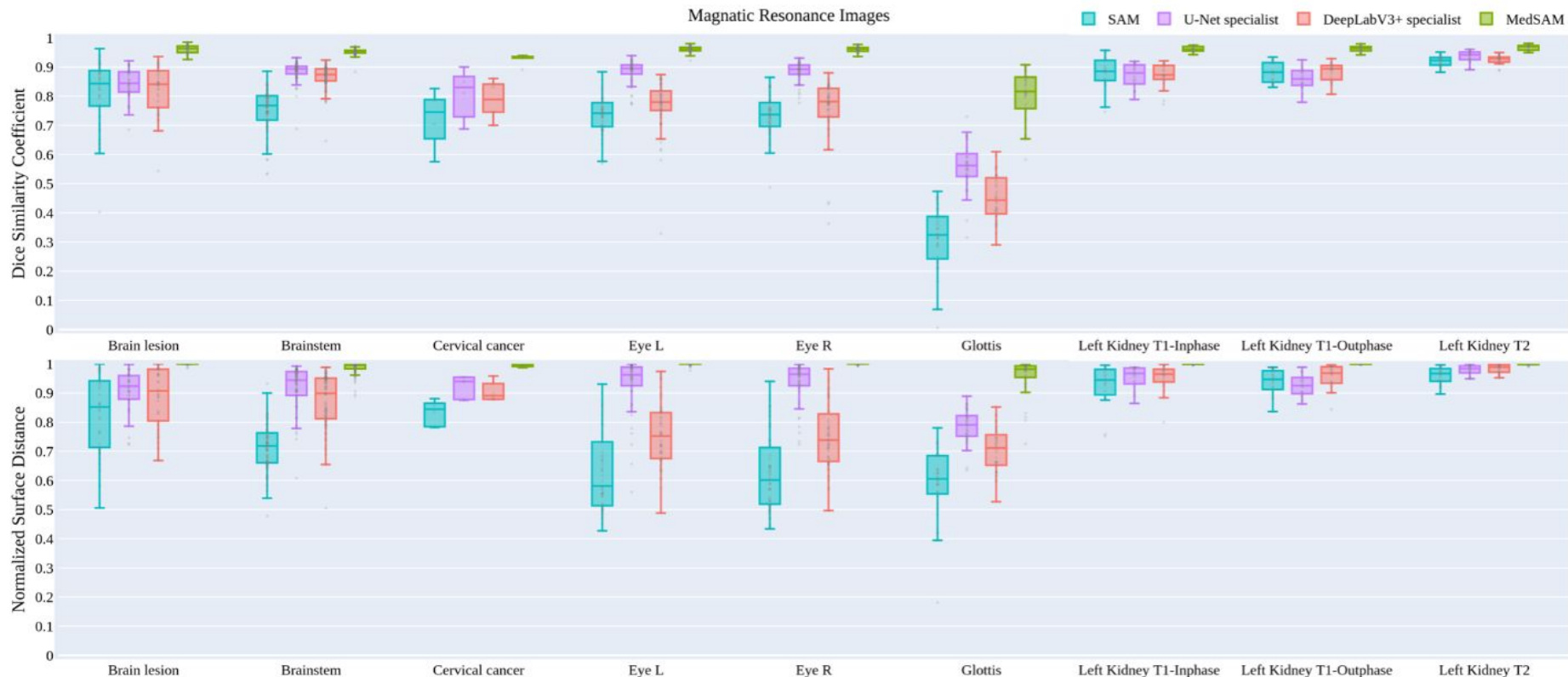# Some external validation on MRIs (DSC and NSD)

Fig. 11. Box plots of dice similarity coefficient and normalized surface distance scores for each MR segmentation task in external validation. The center line within the box represents the median value, with the bottom and top bounds of the box delineating the 25th and 75th percentiles, respectively. Whiskers are chosen to show the 1.5 of the interquartile range. Source data are provided as a Source Data file.

Discussion

Limitations:

- **Modality imbalance in the training set**: underperform on less-represented modalities

- **Bounding box prompt can be ambiguous** for segmentation of vessel-like branching structures

Conclusion:

- Holds great potential to accelerate the advancement of new diagnostic and therapeutic tools

Thank you for your attention

Context: Partial volume effect VS segmentation

Context: Partial volume effect VS segmentation

Discussion