ScribblePrompt: Fast and Flexible Interactive Segmentation for Any Biomedical Image

Hallee E. Wong[1,2], Marianne Rakic[1,2], John Guttag[1], and Adrian V. Dalca[1,2,3]

[1] MIT CSAIL, Cambridge, MA, USA
[2] Martinos Center, Massachusetts General Hospital, Charlestown, MA, USA
[3] Harvard Medical School, Boston, MA, USA
{hallee,mrakic,guttag,adalca}@mit.edu

Table of Contents

A Demo and Code

An interactive demo, code, model weights, and the MedScribble dataset are available at

https://scribbleprompt.csail.mit.edu

B ScribblePrompt Implementation

B.1 Prompt Simulation

In this section, we provide illustrations of the prompt simulation process. Each of these click and scribble simulation algorithms can be applied to the ground truth label (or false negative error region) to simulate positive clicks/scribbles and to the background (or false positive error region) to simulate negative clicks/scribbles.

Scribbles We simulate diverse and varied scribbles by first generating clean scribbles using one of three methods: (i) line scribbles, (ii) centerline scribbles or (iii) contour scribbles. Then, we break up and warp the scribbles to add more variability to account for human error.

Line Scribbles. Fig. 1 illustrates the process of simulating line scribbles.

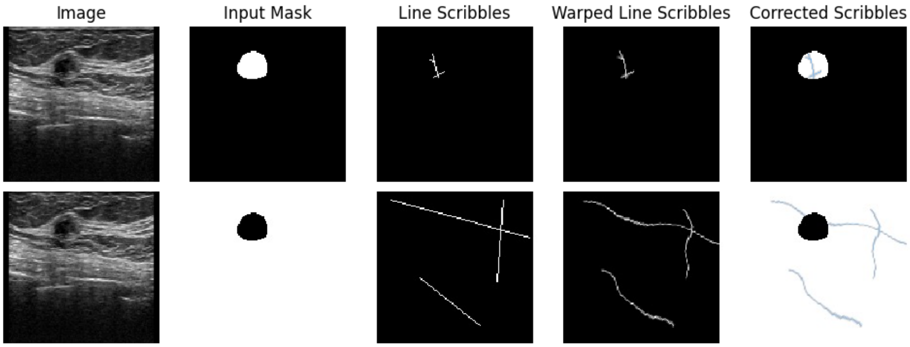


Fig. 1: Line scribbles. Given an input mask z, we draw random lines by sampling two end points from $\{(u,v)|z_{uv}=1\}$. We use a random deformation field to warp the line scribbles and then multiply by the binary input mask z to correct parts of the scribble that were warped outside the mask. We can simulate positive scribbles by applying the algorithm to the ground truth label y (top) and negative scribbles by applying the algorithm to the background $1-y$ (bottom).

Centerline Scribbles. Fig. 2 illustrates the process of simulating centerline scribbles.

Contour Scribbles. Fig. 3 illustrates the process of simulating contour scribbles.

Interior Border Region Clicks. Fig. 4 illustrates the process for simulating interior border region clicks.

B.2 Architecture and Training

We discuss some of the modeling decisions in ScribblePrompt-UNet and ScribblePrompt-SAM.

Normalization Layers. In preliminary experiments, we evaluated normalization layers in the ScribblePrompt-UNet architecture such as Batch Norm [37], Instance Norm [90], Layer Norm [5], and Channel Norm [93]. Including normalization did not improve the mean Dice on validation data compared to using no normalization layers (Fig. 5).

Loss Function. In preliminary experiments, we trained ScribblePrompt with Soft Dice Loss [21], a combination of Soft Dice Loss and Binary Cross-Entropy

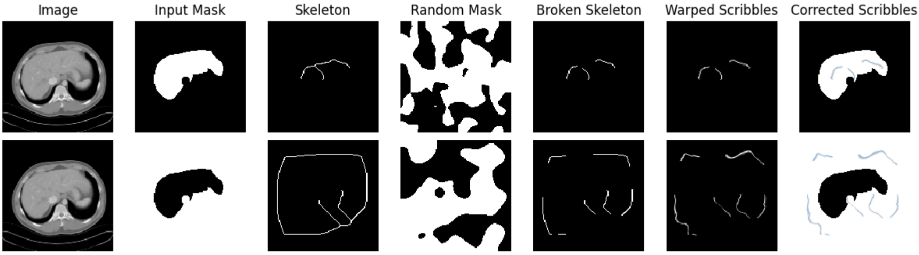


Fig. 2: Centerline scribbles. Given an input mask, we apply a thinning algorithm [96] to get a 1-pixel wide skeleton. We break up the skeleton using a random mask and use a random deformation field to warp the broken skeleton. Lastly, we multiply the scribble mask by the input binary mask to remove parts of the scribble that were warped outside the input mask. We can simulate positive scribbles by applying the algorithm to the label y (top) and negative scribbles by applying the algorithm to the background $1 - y$ (bottom).

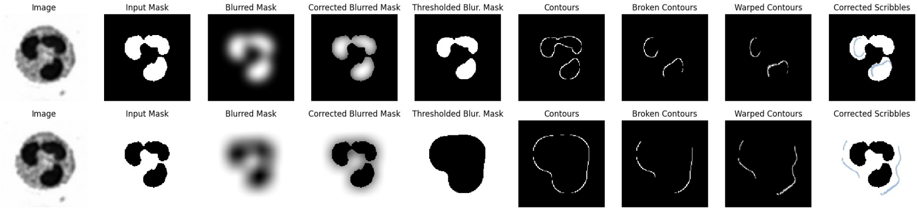


Fig. 3: Contour scribbles. We simulate a rough contour of the desired segmentation within the boundaries of the label. Given a mask z, We first blur the mask to reduce the size of the label such that $\tilde{z} = \min(z, z \circ G_k)$, where G_k is a Gaussian blur kernel. Then we apply a threshold $\tilde{z} < h$ sampled in some intensity range $h \sim U[\tilde{z}_{min}, \tilde{z}_{max}]$ and extract a contour inside the boundary of the mask. We break up the contour using a random mask and use a random deformation field to warp the broken contour. Lastly, we multiply the scribble mask by the input binary mask to correct parts of the scribble that were warped outside the mask. We can simulate positive scribbles by applying the algorithm to the label y (bottom) and negative scribbles by applying the algorithm to the background $1 - y$ (top).

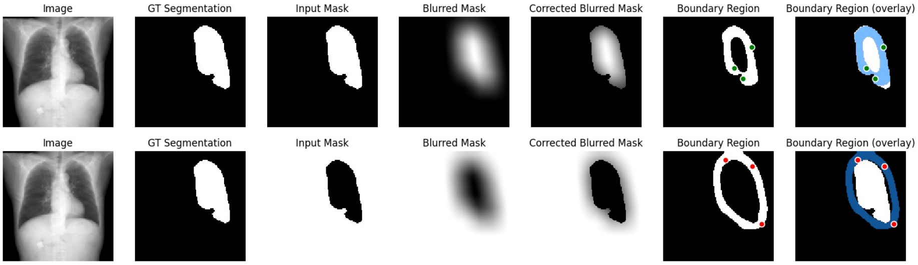


Fig. 4: Interior border region clicks. We sample clicks from a border region inside the boundary of a given mask. Given a mask z, we first blur the mask to reduce the size of the label such that $\tilde{z} = \min(z, z \circ G_k)$ where G_k is a Gaussian blur kernel. We then sample click coordinates from $\{(u,v) | \tilde{z}_{uv} \in [a,b]\}$, where $a, b \sim U[\tilde{z}_{min}, \tilde{z}_{max})$ are thresholds sampled in some intensity range. We show the simulation process for negative border region clicks on the background $1-y$ (top) and positive border region clicks on the label y (bottom).

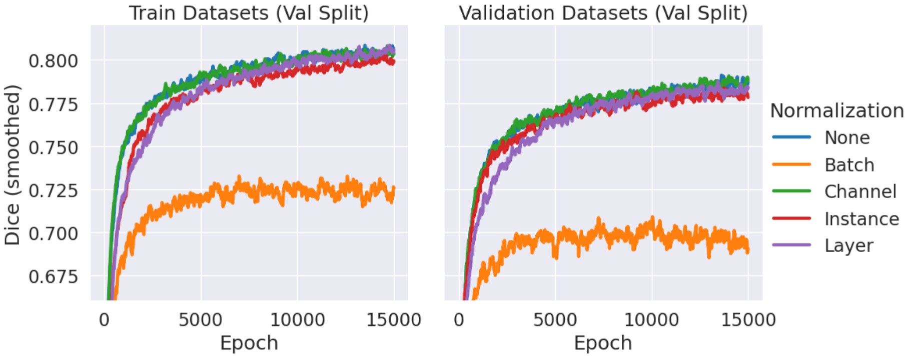


Fig. 5: Training ScribblePrompt-UNet with different normalization layers. We show mean Dice averaged across five iterative predictions (using the training procedure for simulating interactions). At each epoch, we evaluate on 1,000 randomly sampled examples from the validation splits of the 65 training datasets and validation splits of the nine validation datasets. Dice was smoothed using Exponential Weighted Mean with $\alpha = 0.1$.

Loss, and a combination of Soft Dice Loss and Focal Loss [57], similar to [45]. In the latter two losses, Dice Loss and BCE Loss or Focal Loss are weighted equally. We found that the combination of Soft Dice Loss and Focal Loss resulted in slightly higher mean Dice on the validation data for ScribblePrompt-UNet and ScribblePrompt-SAM. Fig. 6 shows Dice recorded during training in preliminary experiments with ScribblePrompt-UNet.

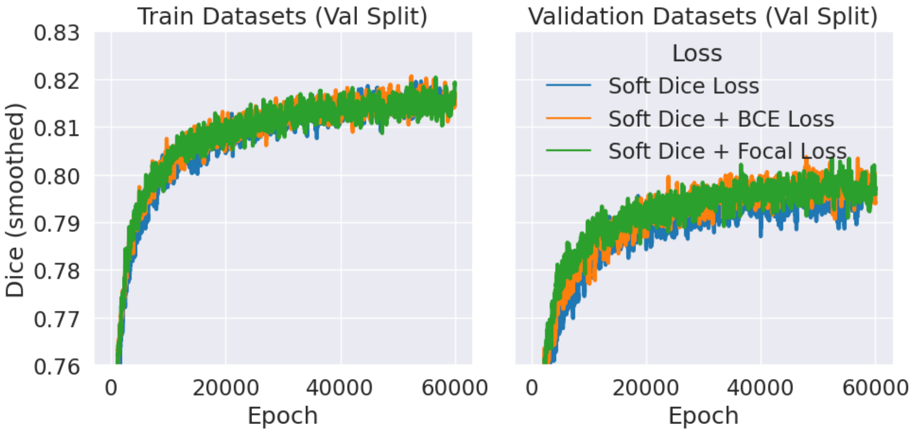


Fig. 6: Training ScribblePrompt-UNet with different loss functions. We report Dice averaged across five iterative predictions (using the training procedure for simulating interactions). At each epoch, we evaluate on 1,000 randomly sampled examples from the validation splits of the 65 training datasets and validation splits of the nine validation datasets. Dice was smoothed using Exponential Weighted Mean with $\alpha = 0.1$.

ScribblePrompt-UNet Inputs. We encode each prompt type in an input channel for ScribblePrompt-UNet. The input to ScribblePrompt-UNet has size $5 \times h \times w$ consisting of the input image x^t, bounding box encoding, positive click/scribble encoding, negative click/scribble encoding, and the logits of the previous prediction $\hat{y}^t_{i-1}$. For the first prediction, we set the previous prediction channel to zeros. We encode bounding boxes in a binary mask that is 1 inside the box(es) and 0 everywhere else. We encode positive and negative clicks using binary masks where a pixel is 1 if has been clicked and 0 otherwise. We encode positive and negative scribbles as masks on $[0, 1]$ and combine them with the masks encoding clicks. Representing the interactions as masks is advantageous because inference time does not scale with the number of interactions.

ScribblePrompt-SAM Details. To train ScribblePrompt-SAM, we took the pre-trained weights from SAM [45] with ViT-b backbone and froze all components of the network except for the decoder.

The SAM architecture can make predictions in single-mask mode or multi-mask mode. In *single-mask mode*, the decoder outputs a single predicted segmentation given an input image and user interactions. In *multi-mask mode*, the decoder predicts three possible segmentations and then outputs the segmentation with the highest predicted IoU by a MLP. We trained and evaluated ScribblePrompt-SAM in multi-mask mode to maximize the expressiveness of the architecture. During training we included a MSE term in the segmentation loss to train the MLP to predict the IoU of the predictions, as in [45].

B.3 Synthetic Labels

To help reduce task overfitting – memorizing the segmentation task for single-label datasets and thus ignoring user prompts – we introduce a mechanism to generate synthetic labels. During training, for a given sample (x_0, y_0), with probability p_{synth} we replace y_0 with a synthetic label y_{synth}.

We use a superpixel algorithm [22] with randomly sampled scale parameter $\lambda \sim U[1, 500]$ to partition the image x_0 into a map of k superpixels, $z \in \{1, \ldots, k\}^{n \times n}$. Then, we randomly select a superpixel $c \sim \mathrm{Cat}(\{1, \ldots, k\}, 1/k)$ as the synthetic label $y_{synth} := \mathbb{1}[z = c]$. Fig. 7 shows examples of training images and the corresponding maps of possible synthetic labels with different λ.

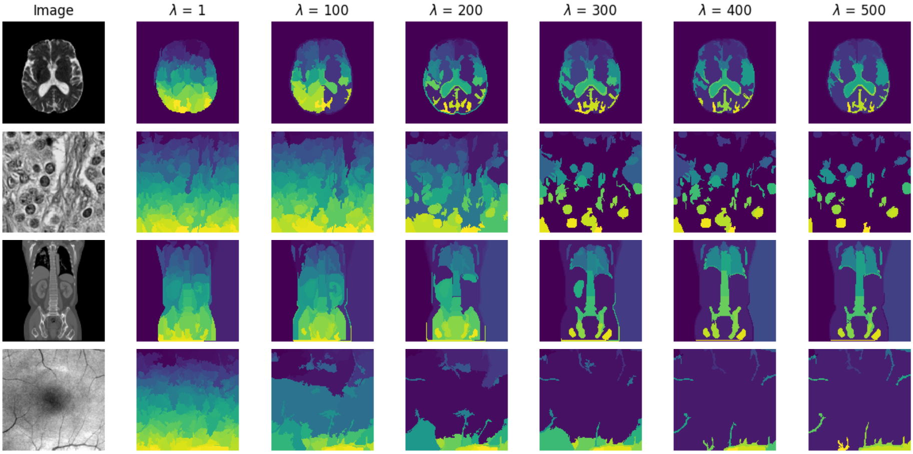


Fig. 7: Examples of possible synthetic labels. Each color in the maps is a different synthetic label. During training, we replace a given label y_0 with a synthetic label y_{synth} with probability p_{synth}. To generate y_{synth}, we apply a superpixel algorithm with randomly sampled scale parameter λ to the image x_0 and then randomly select a superpixel as the synthetic label. We show examples of the synthetic label maps generated using a superpixel algorithm [22] with different λ.

C Data

We build on large dataset gathering efforts like MegaMedical [13,81] to compile a collection of 77 open-access biomedical imaging datasets for training and evaluation, covering over 54k scans, 16 image types, and 711 labels. We gathered datasets with a particular focus on Microscopy, X-Ray, and Ultrasound modalities, which were not as well represented in the original MegaMedical [13]. The full list of datasets is provided in Tab. 2 and Tab. 3.

We define a 2D segmentation task as a combination of (sub)dataset, axis (for 3D modalities), and label. For datasets with multiple segmentation labels, we consider each label separately as a binary segmentation task. For datasets with sub-datasets (e.g., malignant vs. benign lesions) we consider each cohort as a separate task. For multi-annotator datasets, we treat each annotator as a separate label. For instance segmentation datasets, we sampled one instance at a time during training.

For 3D modalities we use the slice with maximum label area ("maxslice") and the middle slice ("midslice") for each volume for training of ScribblePrompt. We report results evaluating on maxslices, but we observed similar trends evaluating on midslices.

Division of Datasets. The division of datasets and subjects for training, model selection, and evaluation is summarized in Tab. 1. The 77 datasets were divided into 65 training datasets (Table 3, 12 evaluation datasets. Data from 9 (out of 12) of the evaluation datasets was used in model development for model selection, and final evaluation. The other 3 evaluation datasets were completely held-out from model development and only used in the final evaluation.

Division of Subjects. We split each of the 77 datasets into 60% train, 20% validation, and 20% test by subject. We used the "train" splits from the 65 training datasets to train ScribblePrompt models. We use the "validation" splits from the 65 training datasets and 9 validation datasets for model selection. We report final evaluation results across 12 evaluation sets consisting of the "test" splits of the 9 validation datasets *and* "test" splits of the 3 test datasets to maximize the diversity of tasks and modalities in our evaluation set (Tab. 1). No data from the 9 validation datasets or 3 test datasets were seen by ScribblePrompt models during training. For TotalSegmentator [92], we only evaluated on 20 examples per task due to the large number of tasks in the dataset. In total, the evaluation data cover 608 segmentation tasks.

Image Processing. We rescale image intensities to [0,1]. For methods using the SAM architecture, we convert the images to RGB and apply the pixel normalization scheme in [45].

Image Resolution. We resized images to 128x128 for training of ScribblePrompt. We used this resolution to reduce training time during model development and to be able to conduct more thorough experiments. The ScribblePrompt approach is not tied to a particular resolution.

We conducted the experiments with MedScribble and simulated interactions with 128^2 size images. For the ACDC scribbles dataset and the user study we evaluated 256^2 size images to test ScribblePrompt's performance at higher resolutions. Although the ScribblePrompt-UNet architecture can take variable size inputs, we found downsizing the image to 128^2 for inference then upsampling the prediction to the input image size produced the highest Dice predictions.

For each method we resize the image to the method's training image size before running inference. Although the SAM architecture takes input images of size 1024^2 (or 256^2 in the case of SAM-Med2D), the the network outputs predictions of size 256^2 that are up-sampled to the input image size. MIDeepSeg takes 96^2 size images as inputs (after automatic cropping) and outputs predictions of size 96^2.

Interactive Baselines. SAM-Med2D used three of our evaluation datasets (ACDC [9], BTCV [50] and TotalSegmentator [92]) as training datasets [94]. MedSAM used two of our evaluation datasets (TotalSegmentator [92] and BUID [4]) as training datasets [66].

Supervised Baselines. We trained fully-supervised baselines for 10 of our evaluation datasets. For those datasets, We used the train and validation splits to train a fully-supervised nnUNet [38] for each 2D task (Tab. 1). We report final results for all methods on the test splits of the evaluation datasets.

Table 1: Dataset split overview. Each dataset was split into 60% train, 20% validation and 20% test by subject. Data from the "train" splits of the 65 training datasets were used to train the models. The ScribblePrompt models did not see any data from the validation datasets or test datasets during training. Data from the "validation" split of the 9 validation datasets was used for ScribblePrompt (SP) model selection and baseline model selection (e.g., single-mask vs. multi-mask mode for SAM). We report final results on 12 "evaluation sets": data from the "test" splits of the 9 validation datasets and the "test" splits of the 3 test datasets. To train the fully-supervised nnUNet baselines, we used the training and validation splits of the 12 evaluation datasets.

		Split within each dataset by subject		
Dataset Group	No. Datasets	Training Split (60%)	Validation Split (20%)	Test Split (20%)
Training Datasts	65	SP training	SP model selection	Not used
Validation Datasets	9	nnUNet training	SP and baselines model selection, nnUNet training	Final evaluation
Test Datasets	3	nnUNet training	nnUNet training	Final evaluation

Table 2: Validation and test datasets. We assembled the following set of datasets to evaluate ScribblePrompt and baseline methods. For the relative size of datasets, we include the number of unique scans (subject and modality pairs) that each dataset has. These datasets were unseen by ScribblePrompt during training. Three test datasets were completely held-out from model selection and development. The validation splits of the other 9 (validation) datasets were used for model selection. We report final results on the test splits of these 12 datasets.

Dataset Name	**Description**	**Scans**	**Labels**	**Modalities**
ACDC [9]	Left and right ventricular endocardium	99	3	cine-MRI
BTCV Cervix [50]	Bladder, uterus, rectum, small bowel	30	4	CT
BUID [4]	Breast tumors	647	2	Ultrasound
COBRE [3,17,23]	Brain anatomy	258	45	T1-weighted MRI
DRIVE [89]	Blood vessels in retinal images	20	1	Optical camera
HipXRay [32]	Ilium and femur	140	2	X-Ray
PanDental [1]	Mandible and teeth	215	2	X-Ray
SCD [80]	Sunnybrook Cardiac Multi-Dataset Collection	100	1	cine-MRI
SCR [26]	Lungs, heart, and clavicles	247	5	X-Ray
SpineWeb [99]	Vertebrae	15	1	T2-weighted MRI
TotalSegmentator [92]	104 anatomic structures (27 organs, 59 bones, 10 muscles, and 8 vessels)	1,204	104	CT
WBC [100]	White blood cell cytoplasm and nucleus	400	2	Microscopy

Table 3: Training datasets. We assembled the following set of datasets to train ScribblePrompt. For the relative size of datasets, we have included the number of unique scans (subject and modality pairs) that each dataset has.

Dataset Name	**Description**	**Scans**	**Modalities**
AbdominalUS [91]	Abdominal organ segmentation	1,543	Ultrasound
AMOS [39]	Abdominal organ segmentation	240	CT, MRI
BBBC003 [60]	Mouse embryos	15	Microscopy
BBBC038 [14]	Nuclei instance segmentation	670	Microscopy
BrainDev [29, 30, 48, 85]	Adult and neonatal brain atlases	53	Multimodal MRI
BrainMetShare [31]	Brain tumors	420	Multimodal MRI
BRATS [6, 7, 74]	Brain tumors	6,096	Multimodal MRI
BTCV Abdominal [50]	13 abdominal organs	30	CT
BUSIS [97]	Breast tumors	163	Ultrasound
CAMUS [51]	Four-chamber and Apical two-chamber heart	500	Ultrasound
CDemris [40]	Human left atrial wall	60	CMR
CHAOS [41, 42]	Abdominal organs (liver, kidneys, spleen)	40	CT, T2-weighted MRI
CheXplanation [83]	Chest X-Ray observations	170	X-Ray
CT2US [88]	Liver segmentation in synthetic ultrasound	4,586	Ultrasound
CT-ORG [82]	Abdominal organ segmentation (overlap with LiTS)	140	CT
DDTI [77]	Thyroid segmentation	472	Ultrasound
EOphtha [19]	Eye microaneurysms and diabetic retinopathy	102	Optical camera
FeTA [76]	Fetal brain structures	80	Fetal MRI
FetoPlac [8]	Placenta vessel	6	Fetoscopic optical camera
FLARE [67]	Abdominal organs (liver, kidney, spleen, pancreas)	361	CT
HaN-Seg [78]	Head and neck organs at risk	84	CT, T1-weighted MRI
HMC-QU [20, 44]	4-chamber (A4C) and apical 2-chamber (A2C) left wall	292	Ultrasound
I2CVB [52]	Prostate (peripheral zone, central gland)	19	T2-weighted MRI
IDRID [79]	Diabetic retinopathy	54	Optical camera
ISBI-EM [15]	Neuronal structures in electron microscopy	30	Microscopy
ISIC [16]	Demoscopic lesions	2,000	Dermatology
ISLES [34]	Ischemic stroke lesion	180	Multimodal MRI
KiTS [33]	Kidney and kidney tumor	210	CT
LGGFlair [12, 72]	TCIA lower-grade glioma brain tumor	110	MRI
LiTS [10]	Liver tumor	131	CT
LUNA [86]	Lungs	888	CT
MCIC [27]	Multi-site brain regions of schizophrenic patients	390	T1-weighted MRI
MMOTU [98]	Ovarian tumors	1,140	Ultrasound
MSD [87]	Large-scale collection of 10 medical segmentation datasets	3,225	CT, Multimodal MRI
MuscleUS [71]	Muscle segmentation (biceps and lower leg)	8,169	Ultrasound
NCI-ISBI [11]	Prostate	30	T2-weighted MRI
NerveUS [75]	Nerve segmentation	5,635	Ultrasound
OASIS [35, 69]	Brain anatomy	414	T1-weighted MRI
OCTA500 [53]	Retinal vascular	500	OCT/OCTA
PanNuke [24]	Nuclei instance segmentation	7,901	Microscopy
PAXRay [84]	92 labels covering lungs, mediastinum, bones, and sub-diaphram in Chest X-Ray	852	X-Ray
PROMISE12 [58]	Prostate	37	T2-weighted MRI
PPMI [18, 70]	Brain regions of Parkinson patients	1,130	T1-weighted MRI
QUBIQ [73]	Collection of 4 multi-annotator datasets (brain, kidney, pancreas and prostate)	209	T1-weighted MRI, Multimodal MRI, CT
ROSE [68]	Retinal vessel	117	OCT/OCTA
SegTHOR [49]	Thoracic organs (heart, trachea, esophagus)	40	CT
SegThy [46]	Thyroid and neck segmentation	532	MRI, Ultrasound
ssTEM [25]	Neuron membranes, mitochondria, synapses and extracellular space	20	Microscopy
STARE [36]	Blood vessels in retinal images (multi-annotator)	20	Optical camera
ToothSeg [62]	Individual teeth	598	X-Ray
VerSe [61]	Individual vertebrae	55	CT
WMH [47]	White matter hyper-intensities	60	Multimodal MRI
WORD [64]	Abdominal organ segmentation	120	CT

D Experimental Setup

Training. We use the Adam optimizer [43] and train with a learning rate of 0.0001 until convergence. We use a batch size of 8 for ScribblePrompt-UNet. For ScribblePrompt-SAM we use a batch size of 1, because of memory constraints.

Task Diversity. The final ScribblePrompt-UNet and ScribblePrompt-SAM models were trained with $p_{synth} = 0.5$. Tab. 4 shows the data augmentations we used, similar to the in-task augmentations from [13, 81].

Table 4: Data augmentations during training. For each example, an augmentation is sampled with probability p. We apply augmentations after (optional) synthetic label generation and before simulating user interactions.

Augmentation	p	Parameters
Random Affine	0.5	degrees $\in [0, 360]$ translation $\in [0, 0.2]$ scale $\in [0.8, 1.1]$
Brightness Contrast	0.5	brightness $\in [-0.1, 0.1]$ contrast $\in [0.8, 1.2]$
Gaussian Blur	0.5	$\sigma \in [0.1, 1.1]$ $k = 5$
Gaussian Noise	0.5	$\mu \in [0, 0.05]$ $\sigma \in [0, 0.05]$
Elastic Transform	0.25	$\alpha \in [1, 2]$ $\sigma \in [6, 8]$
Sharpness	0.5	sharpness $= 5$
Horizontal Flip	0.5	None
Vertical Flip	0.5	None

SAM Baselines. For baseline methods using the SAM architecture, we evaluate the models in both "single mask" and "multi-mask" mode. For each baseline method and interaction procedure, we selected the best performing mode based on the average Dice across the validation data and report final results on test data using that mode. In the results with simulated clicks and scribbles by dataset in Appendix F.2, we show results using both modes. For ScribblePrompt-SAM and SAM-Med2D with adapter layers, multi-mask mode resulted in the highest Dice. For SAM-Med2D without adapter layers, we found multi-mask mode led to higher Dice for scribble inputs while single-mask mode led to higher Dice with click inputs. For SAM (ViT-b and ViT-h) and MedSAM, single-mask mode resulted in the higher Dice on average.

E Manual Scribbles

We provide additional setup details and visualizations for the manual scribbles evaluation in Sec. 5.1.

E.1 Setup

MedScribble Dataset. We collected a diverse dataset of manual scribble annotations, which is available at https://scribbleprompt.csail.mit.edu/data. The MedScribble dataset contains annotations from 3 annotators for 64 image segmentation pairs. The examples were randomly selected from the validation split of 14 different datasets (7 training datasets and 7 validation datasets) [1, 9, 32, 35, 36, 41, 42, 50, 51, 53, 69, 80, 84, 91, 99, 100].

For each task, the annotators were shown 5 training examples with the ground truth segmentation and instructed to draw positive scribbles on the region of interest and negative scribbles on the background for 3-5 new images (without seeing the ground truth segmentation). We collected the scribbles using a web app developed in Python using the Gradio library [2]. Two of the annotators used an iPad with stylus and one annotator used a laptop trackpad, to draw the scribbles.

For the manual scribbles evaluation, we report results on a subset of MedScribble, containing only examples from datasets unseen by ScribblePrompt during training. This subset contains 31 image-segmentation pairs (each with 3 sets annotations) covering 7 segmentations tasks from 7 different validation datasets [1, 9, 32, 50, 80, 99, 100]. The subset includes cardiac MRI, dental X-Ray, abdominal organ, spine vertebrae, and cell microscopy segmentation tasks.

ACDC Scribbles Dataset. Like the other datasets we used, we split the ACDC dataset [9] into 60% train, 20% validation and 20% test by subject. We used the validation split for model selection for baseline methods (*e.g.* single-mask vs. multi-mask mode for methods using the SAM architecture). We report results averaged across three labels on all slices for the test subjects.

MedSAM. We only evaluate MedSAM using bounding box prompts because it was fine-tuned exclusively with bounding box prompts and performs poorly with point inputs (Fig. 16). We prompted MedSAM using a bounding box fit to the positive scribbles. For each dataset, we experimented with using the minimum enclosing bounding box or enlarging the box by 5 pixels in each direction and selected the settings that maximized Dice on the validation data. Using the minimum bounding box resulted in higher Dice scores for MedScribble and enlarging the bounding box resulted in in higher Dice scores for ACDC.

SAM. For methods using the SAM architecture (besides MedSAM), we converted the scribble masks to sets of positive and negative clicks for every non-zero pixel in the scribble masks.

ScribblePrompt-UNet. For ScribblePrompt-UNet we found that blurring the scribble masks with a 3x3 Gaussian blur kernel with $\sigma = 0.5$ prior to inference improved Dice scores, perhaps due to differences in the distribution of pixel values between the manually-collected scribbles and simulated scribbles during training. We also experimented with blurring the scribbles for ScribblePrompt-SAM and each of the baseline methods but it did not improve the Dice scores for any other methods.

E.2 Results

Visualizations. Fig. 8 shows predictions for each method using examples from the MedScribble dataset. Fig. 9 shows examples from the ACDC scribbles dataset.

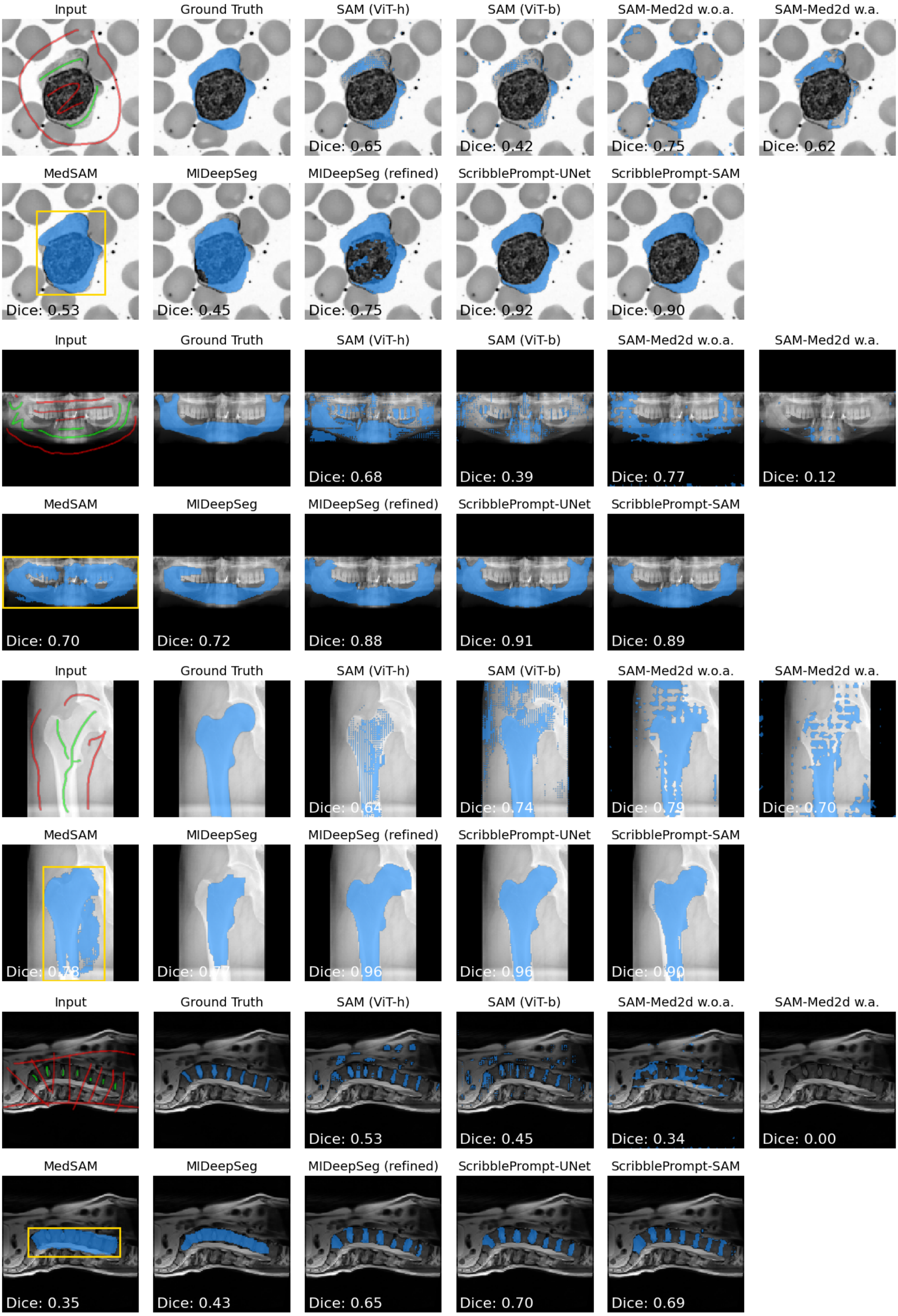


Fig. 8: Example predictions from MedScribble manual scribbles. We evaluate on four examples from the MedScribble dataset. For each method, we show the predicted segmentation given a set of manually-collected positive and negative scribbles as input. For MedSAM, we use a bounding box fit to the positive scribbles as the input.

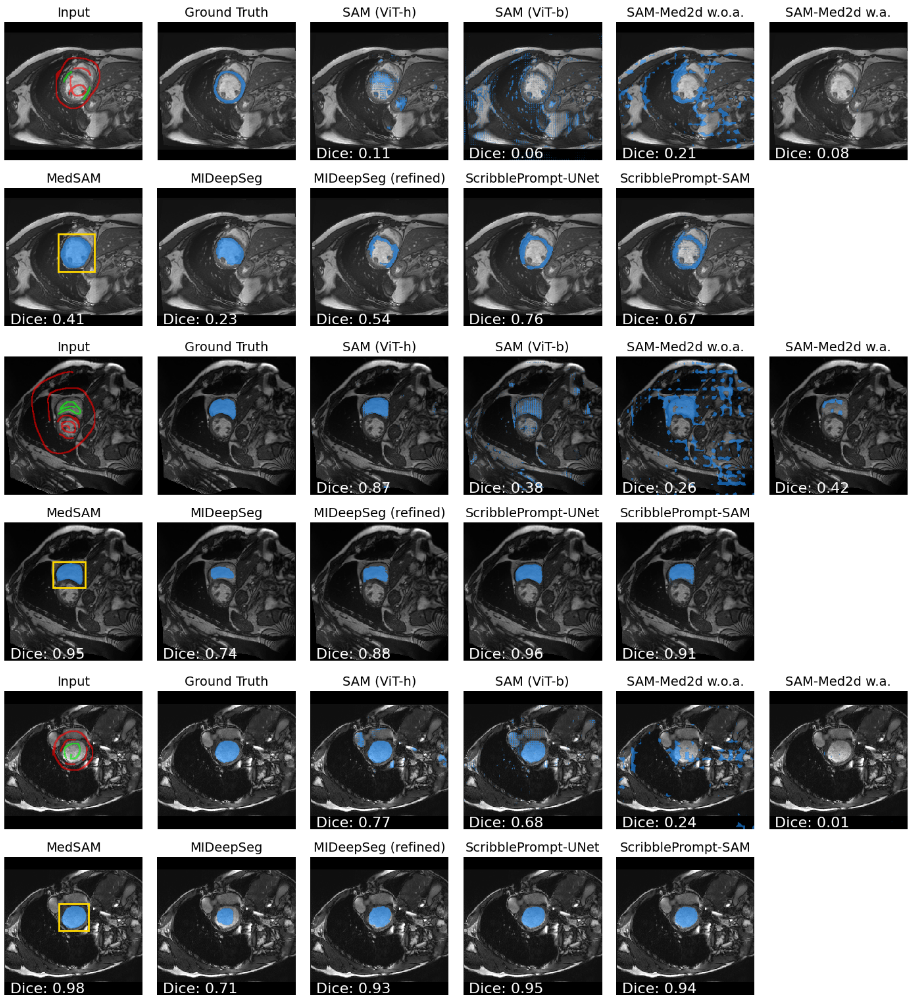


Fig. 9: Example predictions from ACDC manual scribbles. We show examples for each label from the ACDC scribbles dataset [9]. For each method, we show the predicted segmentation given a set of manually-collected positive and negative scribbles as input. For MedSAM, we use a bounding box fit to the positive scribbles with 5 pixels added to each dimension as the input. Scribble thickenss is enlarged for visual clarity.

E.3 Comparison to Scribble-Supervised Learning

We report preliminary results comparing ScribblePrompt to scribble-supervised learning. Scribble-supervised learning methods use scribble annotations as *supervision* to train automatic segmentation models for predicting segmentation given only an input image [28, 54, 56, 63, 95]. These models are task-specific; a new model must be trained using scribble-supervised learning for each new task and training requires many scribble-annotated images from the same task to produce accurate results. In contrast, ScribblePrompt can perform new segmentation tasks at inference time without retraining, using scribbles as *input*.

Setup. We compare ScribblePrompt-UNet to ScribFormer [55], a recent state-of-the-art scribble-supervised learning method, on the ACDC scribbles dataset [9]. Experiments reported in [55] show that ScribFormer's performance varies with the amount of training data, from 0.847 Dice given 14 training subjects to 0.894 Dice given 70 training subjects (and 15 validation subjects) from ACDC.

We evaluate each method given the same test data as in our manual scribbles evaluation: 20 subjects with scribble-annotations for three labels and background. For ScribFormer, we randomly partition the 20 test subjects into 80% train and 20% validation by subject, and train following [55]. We run inference for each model on all 20 test subjects, and report results averaged across the three labels for the 380 slices.

Results. Tab. 5 shows the difference in mean Dice between ScribblePrompt-UNet and ScribFormer is not statistically significant ($p = 0.70$ with a paired t-test). Training ScribFormer required 2 hours using a NVIDIA A100 GPU with 16 CPUs.

Table 5: Comparison to scribble-supervised learning. Mean Dice and HD95 with 95% CI of predicted segmentations for ACDC ($n = 1,140$).

	↑ Dice Score	↓ HD95
ScribFormer	0.85 ± 0.01	4.05 ± 0.99
ScribblePrompt-UNet	0.84 ± 0.01	1.80 ± 0.11

Discussion. Given limited scribble-annotated data from ACDC, ScribblePrompt-UNet predicts segmentations with similar Dice scores and lower HD95 compared to a scribble-supervised learning model trained on the data.

F Simulated Interactions

We present additional results from the experiments in Sec. 5.2 with simulated interactions.

F.1 Bounding Boxes

We evaluate models with simulated bounding box prompts.

Setup. We evaluate segmentation accuracy using Dice score after a single bounding box prompt. We simulate bounding boxes using the same procedure as used was used when training ScribblePrompt: we find the minimum enclosing bounding box for the ground truth label and then enlarge each dimension by $r \sim U[0, 20]$ pixels to account for human error. We exclude MIDeepSeg [65] from this evaluation because it is not designed to make predictions from a single bounding box input.

For methods using the SAM architecture, we apply the pixel normalization scheme in [45] to images before inference. Upon further investigation, MedSAM [66] performed better with images rescaled to $[0, 1]$; we report results for MedSAM with both normalization schemes.

Results. Fig. 10 shows mean Dice after one bounding box prompt. Fig. 11 shows results by dataset. ScribblePrompt-SAM has the highest Dice on average after one bounding box prompt.

Visualizations. Due the ambiguity of many segmentation tasks, its often difficult to predict an accurate segmentation from a single bounding box prompt (Fig. 13). Although ScribblePrompt models produced the highest dice predictions from a single bounding box prompt in Fig. 10, users may not be satisfied with this level of accuracy. Users can still achieve high Dice segmentations with ScribblePrompt by providing additional click and scribble interactions to correct the prediction. We visualize predictions for two examples in Fig. 12 and Fig. 13, after a single bounding box prompt and after correction clicks. MedSAM has the highest mean Dice among the baselines after a single bounding box prompt (Fig. 10), but its usability is limited because it cannot incorporate corrections.

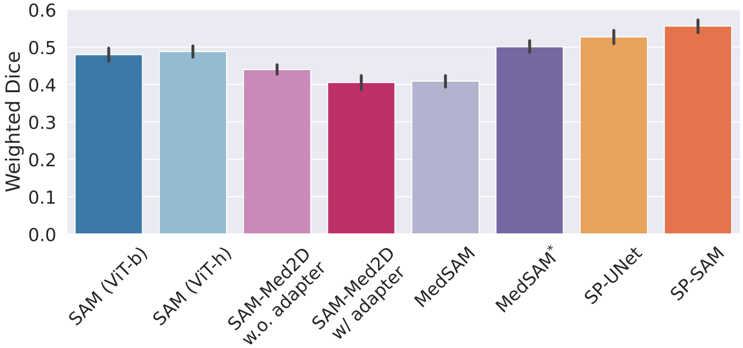


Fig. 10: Results with simulated bounding boxes. Mean Dice on test data from 12 datasets with one simulated bounding box prompt, weighting each dataset equally. SP = ScribblePrompt. MedSAM* indicates MedSAM with input images re-scaled to $[0, 1]$ instead of the pixel normalization from [45]. Errorbars show 95% CI from bootstrapping.

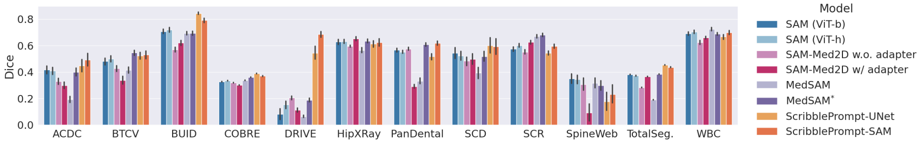


Fig. 11: Results with simulated bounding boxes by dataset. Mean Dice after one simulated bounding box prompt. Among the evaluation datasets, bounding box prompts are the most effective for BUID, a breast ultrasound dataset. MedSAM* indicates MedSAM with input images re-scaled to $[0, 1]$ instead of the pixel normalization from [45]. Errorbars show 95% CI from bootstrapping.

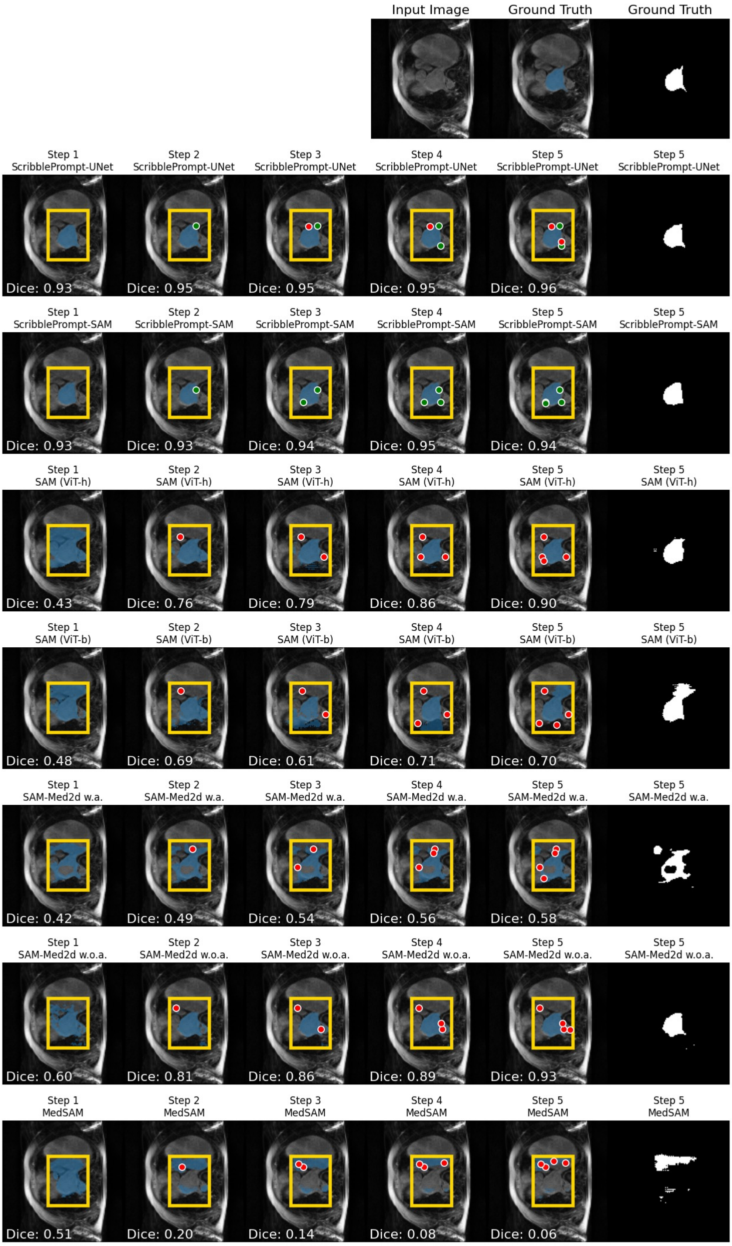


Fig. 12: Bounding box prompt with center correction clicks. We simulate iterative interactive segmentation of the left ventricle in a cardiac MRI from the SCD dataset [80]. This label was seen during training but this dataset was not. ScribblePrompt models produce the highest dice predictions after a single bounding box prompt (first column) and are able to improve their predictions with additional corrections.

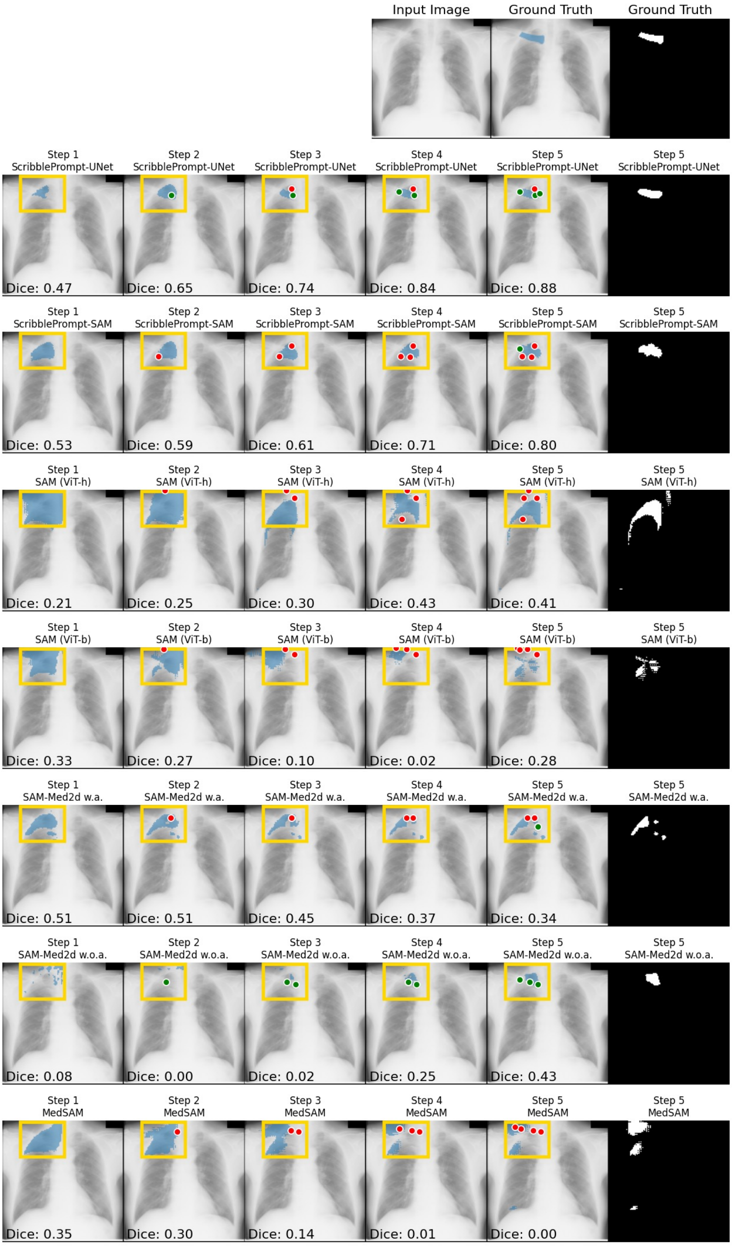


Fig. 13: Bounding box prompt with center correction clicks. We show clavicle segmentation on an frontal chest X-Ray from the SCR dataset [26]. This dataset was completely held-out and this label was unseen during training. None of the methods are able to accurately segment the clavicle from a single bounding box prompt (first column). However, after a few correction clicks, ScribblePrompt-UNet and ScribblePrompt-SAM achieve 0.88 and 0.80 Dice, respectively.

F.2 Scribbles and Clicks

We provide additional setup details, baselines and results for the experiments with simulated scribbles and clicks presented in Sec. 5.2.

Setup. We evaluated each method following three scribble interaction procedures and three click interaction procedures. We provide details below on the MedSAM baseline and additional supervised baselines.

MedSAM. Since MedSAM [66] performs poorly with scribble and click prompts (Fig. 16), we only evaluate it with bounding box prompts. We fit a bounding box to the ground truth segmentation and enlarged each dimension by $r \sim U[0, 10]$ pixels, to match the amount of jitter used during training for MedSAM. We show the mean Dice of segmentations predicted by MedSAM from a single bounding box prompt as a horizontal line (Fig. 14, Fig. 15) because MedSAM cannot incorporate corrections.

Supervised Baselines. We trained fully-supervised task-specific nnUNets [38] for 10 of the evaluation datasets. We show the mean Dice of the segmentations predicted by the ensemble of nnUnets using horizontal lines in the results by dataset (Fig. 22-27).

Results. Fig. 14 shows Dice vs. steps of interaction for three simulated click-focused procedures and three simulated scribble-focused procedures. On average, ScribblePrompt-UNet and ScribblePrompt-SAM have the highest Dice among interactive methods at all steps for all of the simulated interaction procedures. For select interaction procedures we also show HD95 vs. steps of interaction (Fig. 15). ScribblePrompt-UNet and ScribblePrompt-SAM consistently achieve the lowest HD95.

Results by Dataset. Figs. 22, 23, and 24 show quantitative results by dataset for the click-focused interaction procedures. Figs. 25, 26, and 27 show quantitative results by dataset for scribble-focused interaction procedures. ScribblePrompt reaches (or surpasses) fully-supervised nnUNet performance for 5 unseen datasets within 1-3 centerline scribbles steps, and for 10 unseen datasets within 6 scribble steps (Fig. 26).

Visualizations. We show predictions for test examples from evaluation datasets unseen by ScribblePrompt during training. Fig. 17, Fig. 18, and Fig. 19 show iterative predictions from each method using clicks. ScribblePrompt is able to segment large ambiguous objects (Fig. 17), as well as thin structures like vasculature (Fig. 18). For large and complex regions of interest such as white matter in brain MRI (Fig. 19), starting with a few random clicks at once is helpful.

Fig. 20 and Fig. 21 show iterative interactive segmentation with centerline scribbles and line scribbles. ScribblePrompt is able to accurately segment labels unseen during training using scribbles.

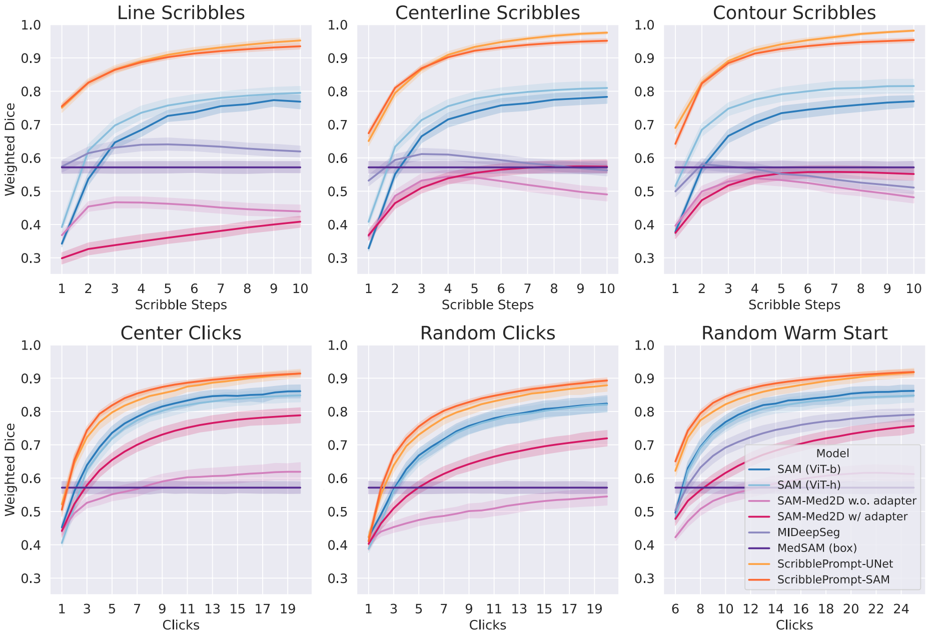


Fig. 14: Dice results with simulated scribbles and clicks. We evaluate methods using three scribble procedures and three click procedures. We measure Dice averaged across twelve evaluation sets (the test splits of the nine validation and three test datasets), weighting each dataset equally. Shaded regions show 95% CI from bootstrapping.

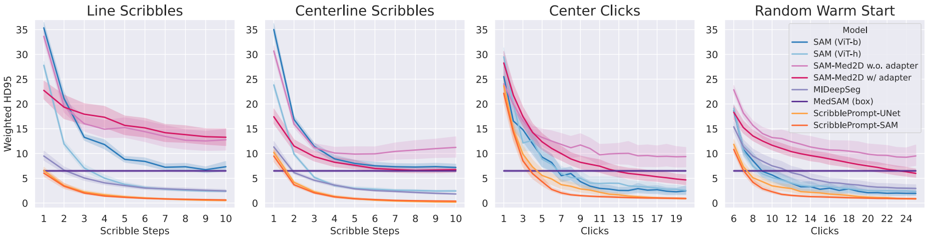


Fig. 15: HD95 results with simulated scribbles and clicks We report HD95 for two scribble procedures and two click procedures. We measure HD95 averaged across twelve evaluation sets (the test splits of the nine validation and three test datasets), weighting each dataset equally. We exclude examples where the ground truth segmentation label was empty or the predicted segmentation was empty. Shaded regions show 95% CI from bootstrapping.

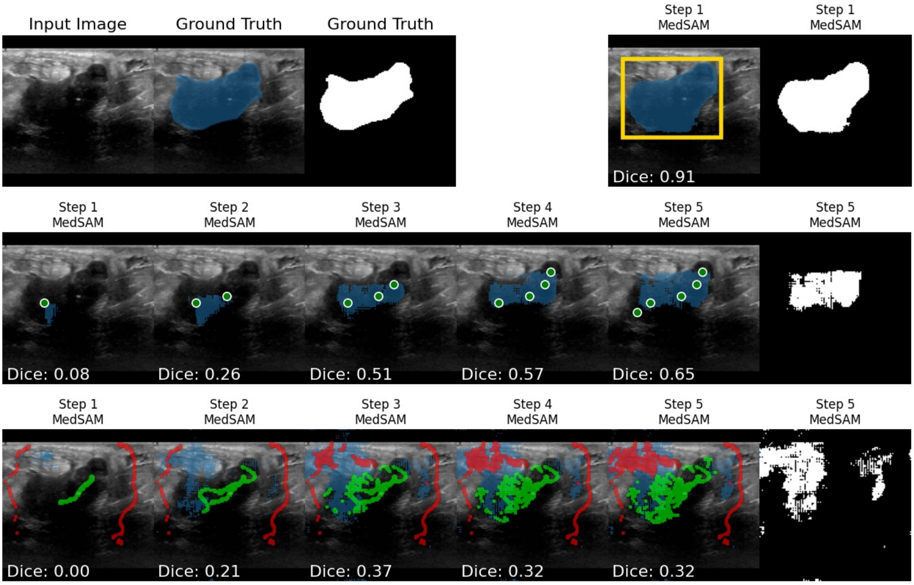


Fig. 16: MedSAM with bounding box, click, and scribble inputs. We do not evaluate MedSAM with click and scribble inputs, which it was not trained for, because it produces poor segmentations with these inputs. Scribble thickness is enlarged for visual clarity.

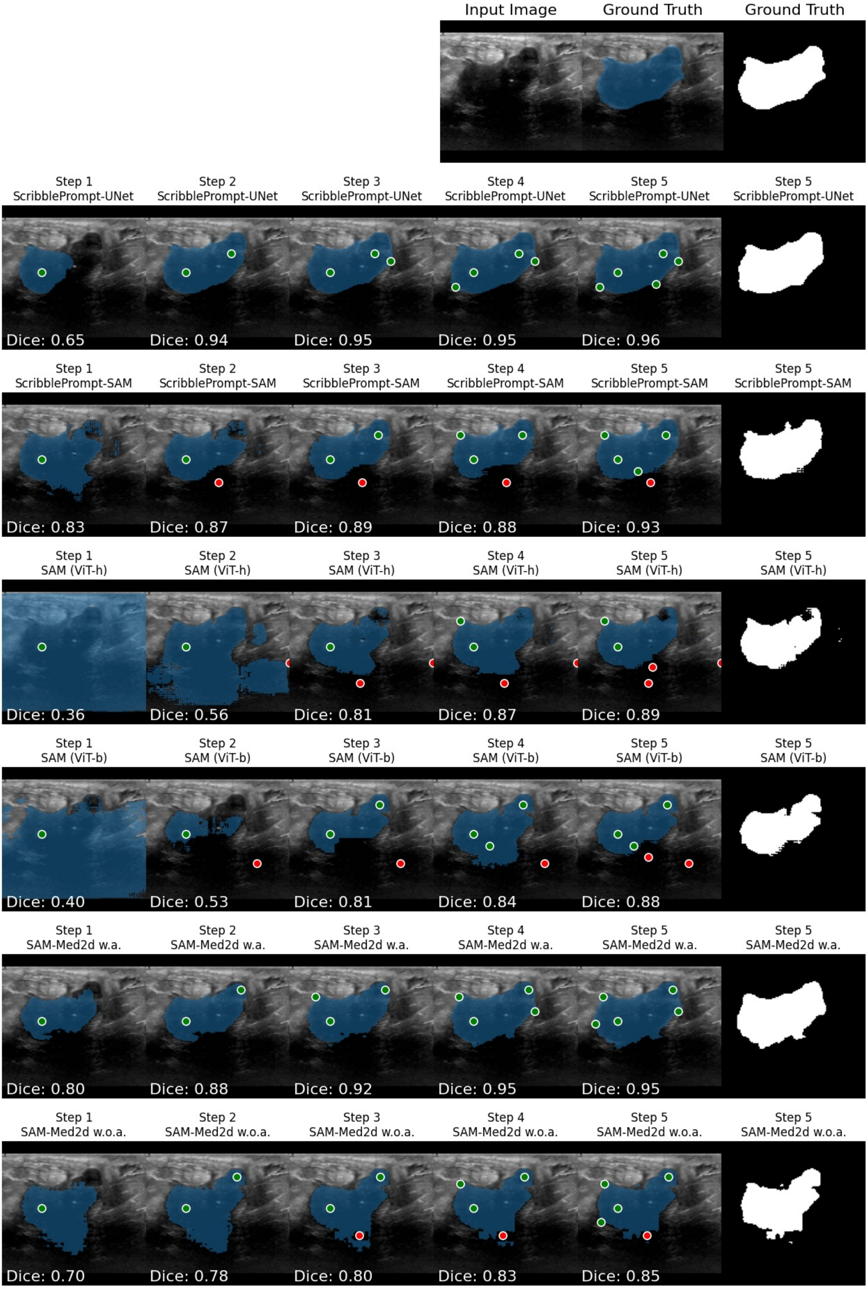
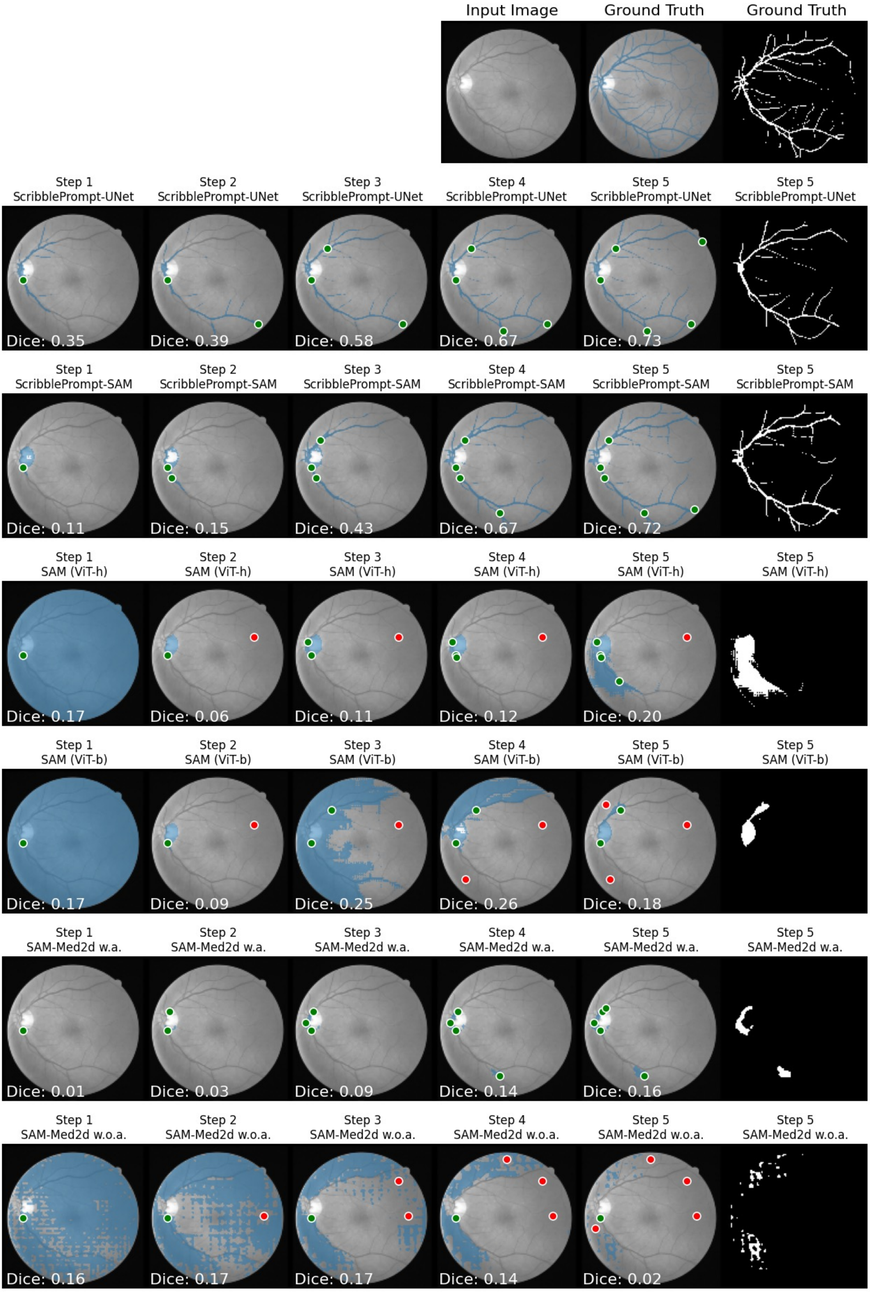


Fig. 17: Example predictions from center clicks. We show an example of interactive segmentation of a malignant tumor in an Ultrasound image from the BUID [4] dataset. This dataset was unseen by ScribblePrompt models during training. We simulate an initial click in the center of the label followed by one correction click in the center of the error at each step.



Input Image
Ground Truth
Ground Truth

Step 1 ScribblePrompt-UNet | Step 2 ScribblePrompt-UNet | Step 3 ScribblePrompt-UNet | Step 4 ScribblePrompt-UNet | Step 5 ScribblePrompt-UNet | Step 5 ScribblePrompt-UNet

Step 1 ScribblePrompt-SAM | Step 2 ScribblePrompt-SAM | Step 3 ScribblePrompt-SAM | Step 4 ScribblePrompt-SAM | Step 5 ScribblePrompt-SAM | Step 5 ScribblePrompt-SAM

Step 1 SAM (ViT-h) | Step 2 SAM (ViT-h) | Step 3 SAM (ViT-h) | Step 4 SAM (ViT-h) | Step 5 SAM (ViT-h) | Step 5 SAM (ViT-h)

Step 1 SAM (ViT-b) | Step 2 SAM (ViT-b) | Step 3 SAM (ViT-b) | Step 4 SAM (ViT-b) | Step 5 SAM (ViT-b) | Step 5 SAM (ViT-b)

Step 1 SAM-Med2d w.a. | Step 2 SAM-Med2d w.a. | Step 3 SAM-Med2d w.a. | Step 4 SAM-Med2d w.a. | Step 5 SAM-Med2d w.a. | Step 5 SAM-Med2d w.a.

Step 1 SAM-Med2d w.o.a. | Step 2 SAM-Med2d w.o.a. | Step 3 SAM-Med2d w.o.a. | Step 4 SAM-Med2d w.o.a. | Step 5 SAM-Med2d w.o.a. | Step 5 SAM-Med2d w.o.a.

Fig. 18: Example predictions from center clicks. We show an example of iterative interactive segmentation of retinal veins in a fundus photograph from the DRIVE dataset [89]. This dataset was unseen by ScribblePrompt models during training. The ScribblePrompt models are able to segment the retinal veins while baselines methods are not able to segment these thin structures.

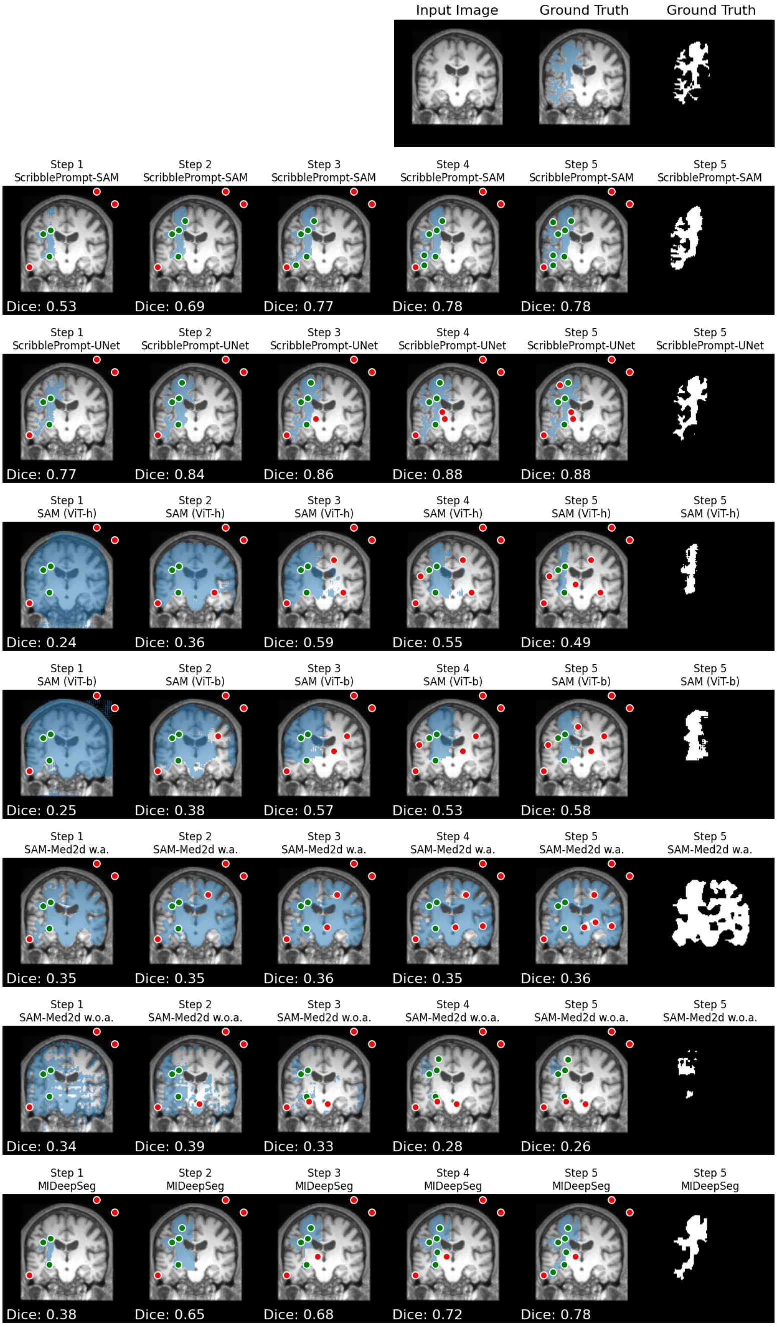


Fig. 19: Example predictions from random clicks and center correction clicks. We show an example of white matter segmentation in a T1 brain MRI from the COBRE dataset [3, 17, 23]. This dataset was completely held-out from ScribblePrompt training and model selection. We simulate interactions following the warm start click protocol: we start with three positive and three negative random clicks, followed by one center correction click per step.

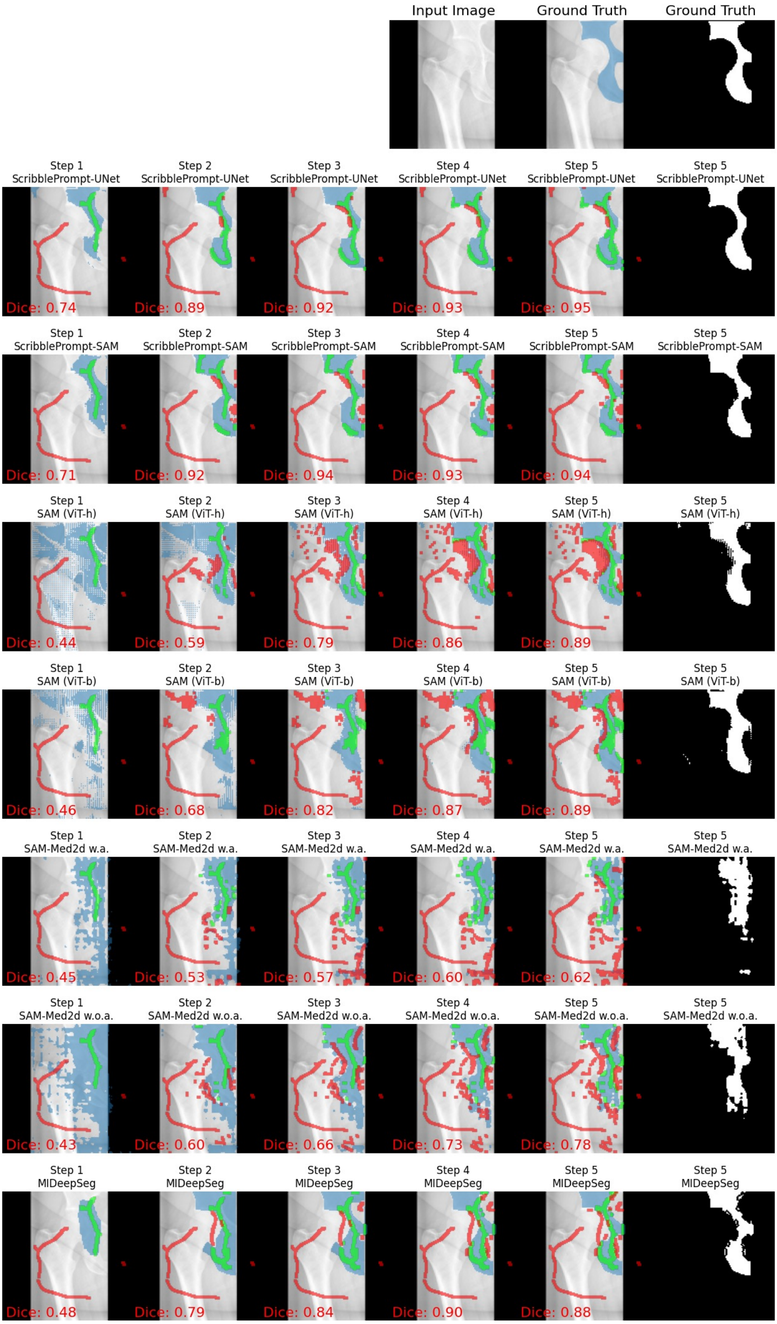


Fig. 20: Example predictions from centerline scribbles. We simulate iterative interactive segmentation of the ilium in an X-Ray from the HipXRay dataset [32]. This dataset, label, and type of X-Ray was not seen by ScribblePrompt models during training. Correction scribbles were simulated separately for each method based on the error region of the previous prediction. ScribblePrompt models have the highest Dice predictions after 5 scribble steps. Scribble thickness is enlarged for visual clarity.

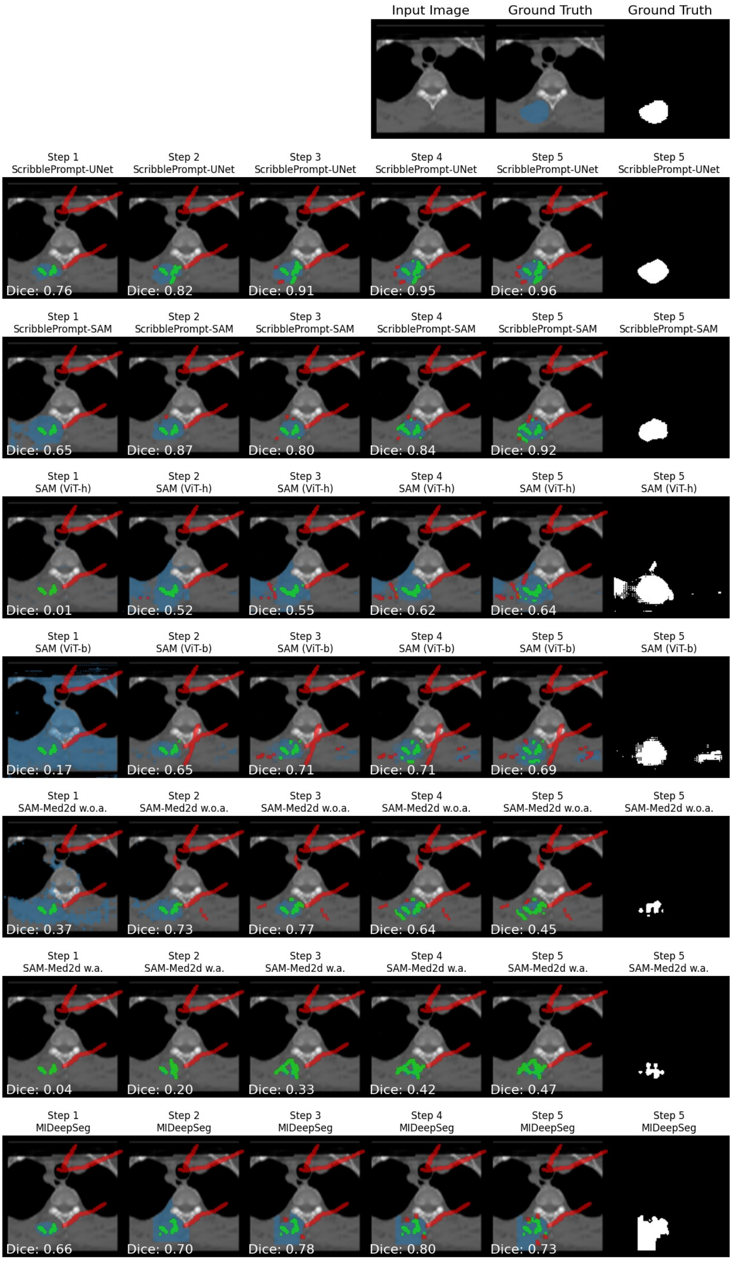


Fig. 21: Example predictions from line scribbles. We simulate iterative interactive segmentation of the left autochthon muscle in a CT from the TotalSegmentator dataset [92]. This dataset was completely held-out and the label was unseen by ScribblePrompt models during training. This segmentation task is challenging because there is little contrast between the region of interest and surrounding tissue. ScribblePrompt models are able to accurately refine their predictions and a achieve Dice ≥ 0.92 after 5 scribble steps. Scribble thickness is enlarged for visual clarity.

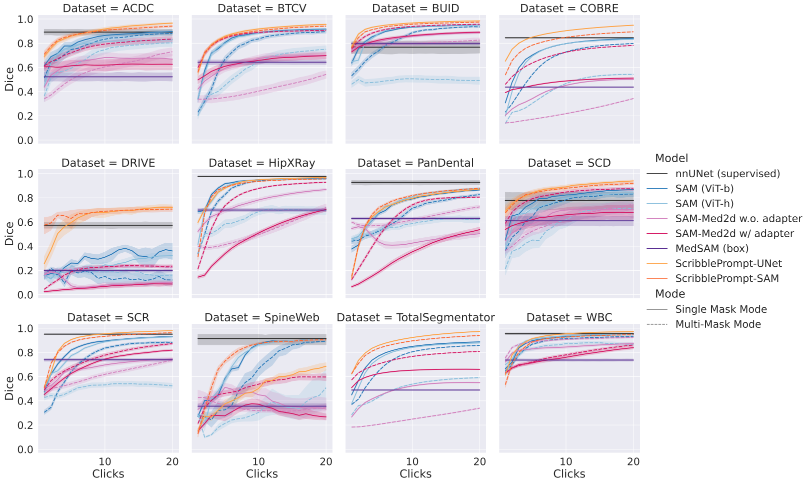


Fig. 22: Results by dataset with center clicks. During the first step, one positive click is placed at the center of the largest component of the ground truth segmentation. In subsequent iterations, one (positive or negative) correction click is placed at the center of the largest component of the error region between the previous prediction and ground truth segmentation.

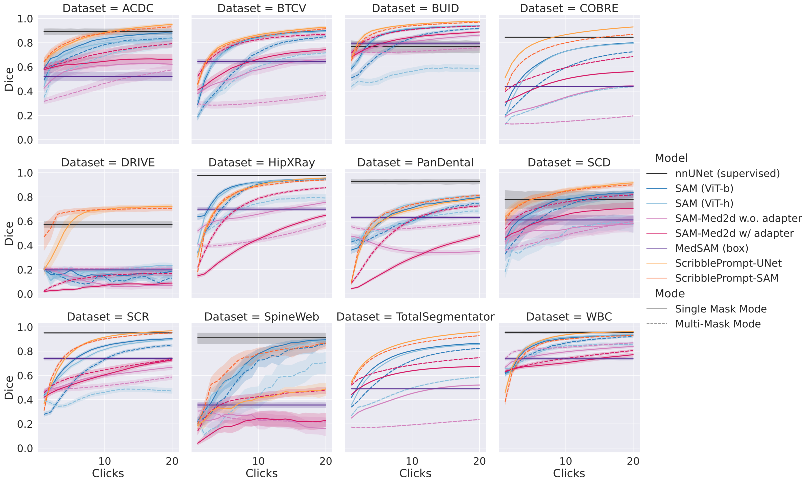


Fig. 23: Results by dataset with random clicks. During the first step, one positive click is randomly sampled from the ground truth segmentation. In subsequent steps, one (positive or negative) correction click is randomly sampled from the error region between the previous prediction and ground truth segmentation.

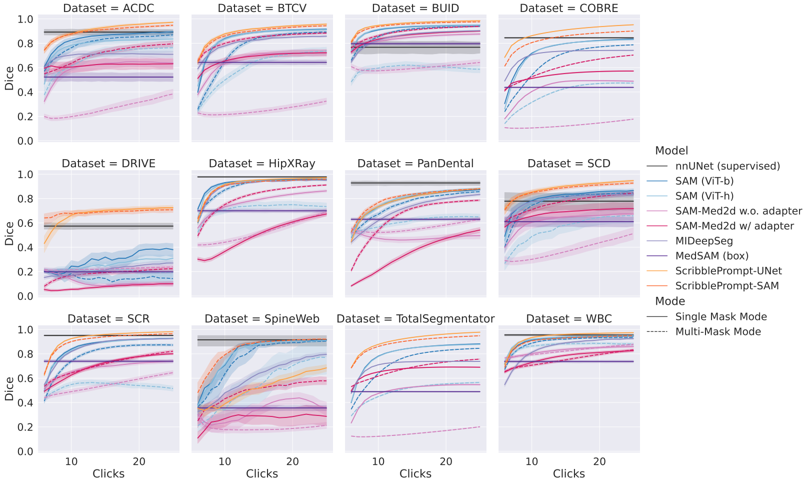


Fig. 24: Results by dataset with random warm start click procedure. During the first step, three positive clicks are randomly sampled from the ground truth segmentation and three negative clicks are randomly sampled from the background. In subsequent steps, one (positive or negative) correction click is placed at the center of the largest component of the error region between the previous prediction and ground truth segmentation.

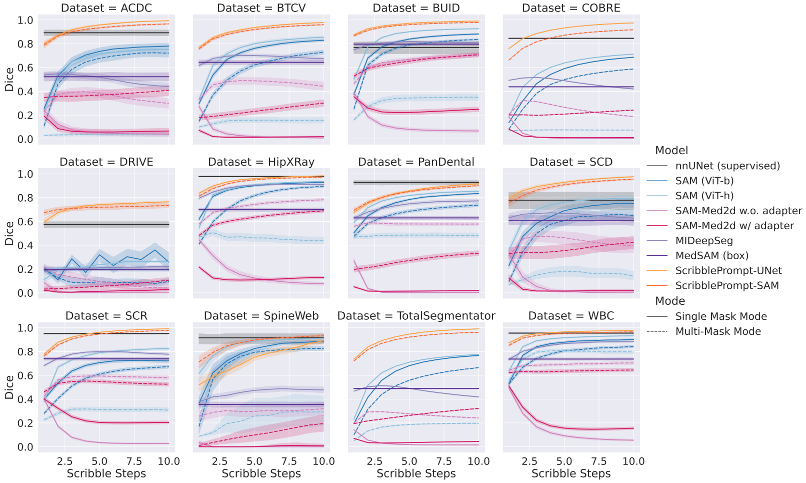


Fig. 25: Results by dataset with line scribbles. During the first step we simulate three positive line scribbles and three negative line scribbles. In subsequent steps, we simulate one (positive or negative) correction line scribble based on the error region between the previous prediction and ground truth segmentation. Each line scribble covers a maximum of 128 pixels.

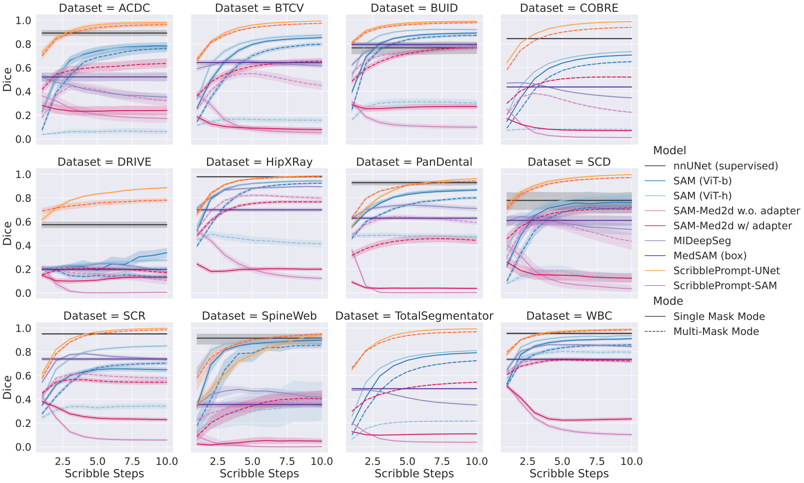


Fig. 26: Results by dataset with centerline scribbles. During the first step, we simulate one positive and one negative centerline scribble. In subsequent steps, we simulate one (positive or negative) correction centerline scribbles based on the error region region between the previous prediction and ground truth segmentation. Each centerline scribble covers a maximum of 128 pixels.

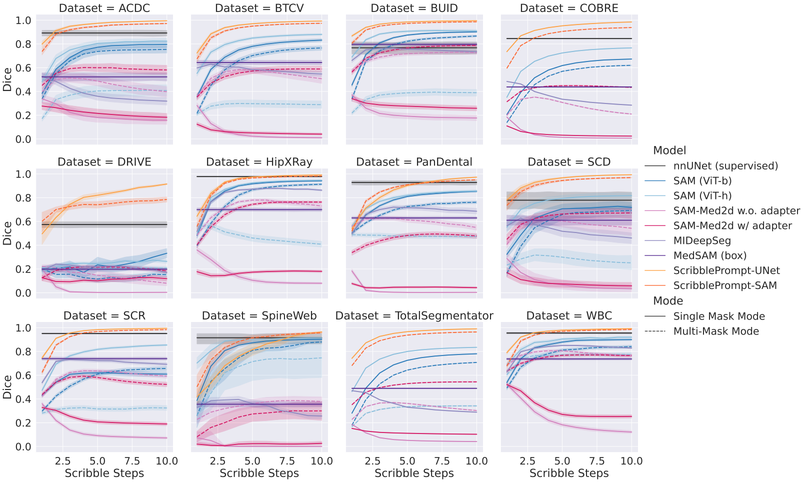


Fig. 27: Results by dataset with contour scribbles. During the first step, we simulate one positive and one negative contour scribble based on the ground truth label. In subsequent steps, we simulate one (positive or negative) correction contour scribble based on the error region region between the previous prediction and ground truth segmentation.Each contour scribble covers a maximum of 128 pixels.

G User Study

We conducted a user study comparing ScribblePrompt-UNet to SAM (ViT-b). We provide additional details on the user study design and implementation.

Study Design. The goal of the user study was to compare ScribblePrompt to the best click-focused baseline method in terms of accuracy (Dice of the final segmentations), efficiency (time to achieve the desired segmentations) and user experience (perceived effort). Participants were given time to familiarize themselves with both models on a fixed set of practice images. Afterwards they used each model to segment a series of nine new test images from nine tasks that were not seen by the model during training (Fig. 28).

The order in which participants used the models, and which image the users were assigned to segment with each model for each task was randomized. We randomly selected one training image per task to include in the set of practice images. We randomly selected two test images per task and randomized the assignment of each image to each model for each participants. Each participant segmented a total of 18 images during the study. The models were also annonymized (i.e., "Model A" and "Model B"). We informed participants that one model was designed to be used with clicks and bounding boxes, while the other was designed for use with clicks, bounding boxes, and scribbles.

For each segmentation task, the participants were shown the target segmentation and were asked to interact with the model until the predicted segmentation closely matched the target or they could no longer improve the prediction. We provided participants with the target segmentation to disentangle the cognitive process of identifying the region of interest from prompting the model to achieve the desired segmentation.

Study Participants. Study participants were neuroimaging researchers at an academic hospital. Although the participants had prior experience with medical image segmentation, they did not necessarily have experience with the specific tasks and types of images used in the study.

We had a total of 29 participants with 16 participants completing all of the segmentations and the exit survey. We observed a higher attrition rate among participants who were assigned to use SAM first, even after being able to freely try out both models during the "practice" phase. Among the 13 participants assigned to use SAM first, 62% did not finish all of their segmentations, compared with 31% among the 16 assigned to use ScribblePrompt first. We report results on the 16 participants who completed all the segmentations and the exit survey.

Implementation. Each participant used a web-based interface powered by a Nvidia Quatro RTX8000 GPU with 4 CPUs. Participants segmented the images at 256 × 256 resolution. The interface was developed in Python using the Gradio library [2]. The interface had a "practice" mode in which users could freely switch between the two models and images from the set of practice images. After experimenting with both models, users clicked a button to begin "recorded activity" mode in which users were led through performing specific segmentation tasks with specific models. Users provided positive/negative scribble inputs, positive/negative click inputs and/or bounding box inputs, and then clicked a button to receive a prediction from the model.

Survey Results. Common factors that influenced participants preference for ScribblePrompt was being able to get accurate predictions from scribbles ("[ScribblePrompt] was more spatially smooth"), the model's responsiveness to a variety of inputs ("it landed on my desired predictions more easily"), and less perceived effort when using the model ("[ScribblePrompt] needed much less guidance"). Participants preferred using clicks and bounding boxes over scribbles with SAM, praising its "snapiness", the effectiveness of "exclusion clicks" and remarking it worked well for "rigid structures". However, participants also noted in some cases "[SAM] did not respect object boundaries", and for tasks such as retinal vein segmentation "[SAM] required lots of clicks and still was not very accurate".

Visualizations. We visualize some of the interactions used by study participants and the resulting predictions in Fig. 28. Study participants used denser clicks when prompting SAM compared to when prompting ScribblePrompt-UNet.

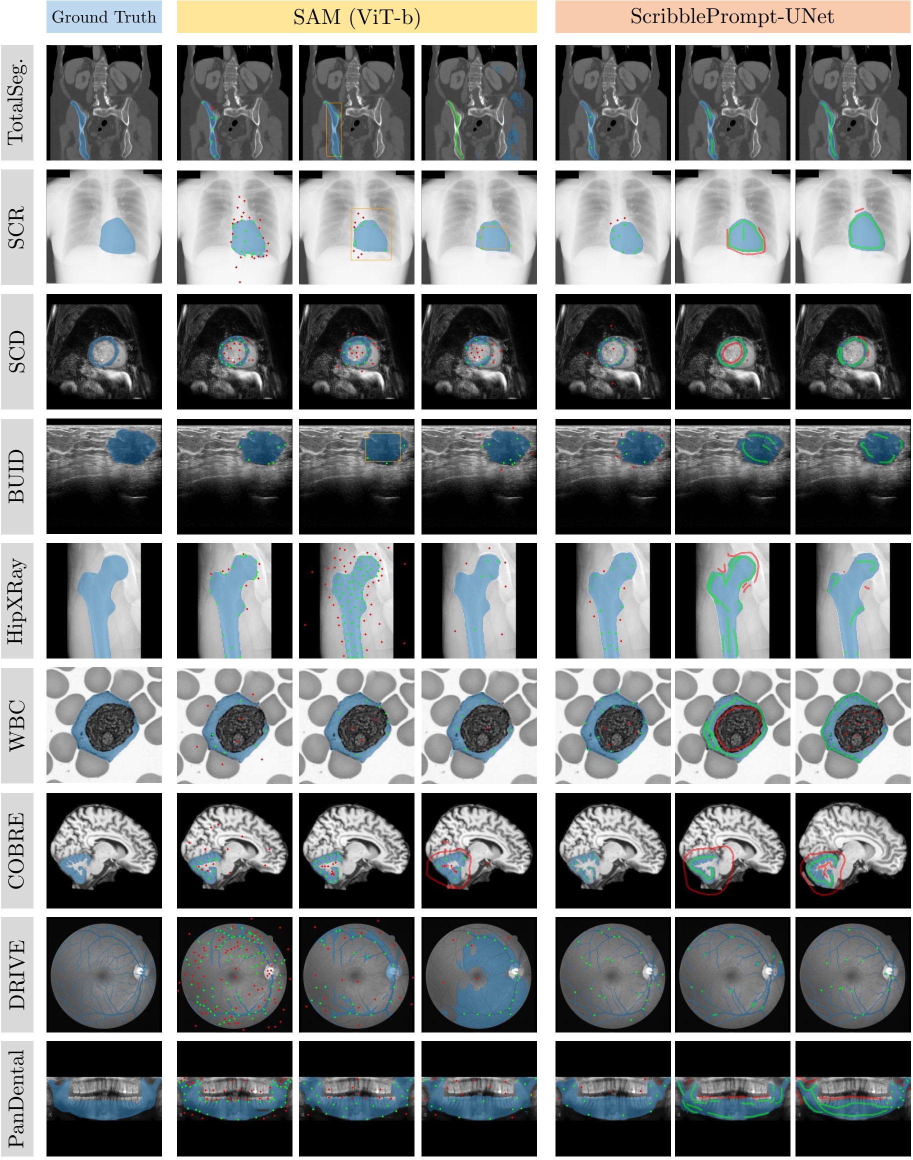


Fig. 28: Example segmentations and interactions from the user study. We show predictions with interactions provided by three study participants for each of the nine segmentation tasks in the user study. For each example, we visualize positive scribble and click inputs in green, negative scribble and click inputs in red, bounding box inputs in yellow, and the predicted segmentation in blue. With SAM, study participants primarily used clicks. With ScribblePrompt, participants used a mix of scribbles and clicks. For the retinal vein segmentation task, participants preferred to use clicks with both models. Participants prompted SAM with denser clicks compared to ScribblePrompt.

H Inference Runtime

Setup. We measure inference time for a random input with a scribble covering 128 pixels. We report mean and standard deviation of inference time across 1,000 runs on a single CPU and on a Nvidia Quatro RTX8000 GPU.

Results. We show performance results in Tab. 6. On a single CPU, ScribblePrompt-UNet requires 0.27 ± 0.04 sec per prediction, enabling the model to be used even in low-resource environments. Prior work on interactive interfaces indicates that < 0.5 sec latency is sufficient for cognitive tasks [59]. ScribblePrompt-UNet is also faster than the baseline methods on a GPU.

Its efficient fully-convolutional architecture gives ScribblePrompt-UNet low latency inference. With SAM, latency scales with the number of interactions because each point is encoded as a 256-dimensional vector embedding. For ScribblePrompt-UNet, clicks and scribbles are encoded in masks, so inference time (per prediction) is constant with the number of interactions.

Table 6: Performance Summary. We measure inference time separately on a single CPU and on an Nvidia Quatro RTX8000 GPU for a prediction with a random scribble input covering 128 pixels. We report mean and standard deviation across 1,000 runs. ScribblePrompt-SAM and MedSAM use the same architecture as SAM ViT-b. Best and second best are highlighted.

Architecture	Param.	CPU Runtime (sec)	GPU Runtime (ms)	GPU Memory
SAM (ViT-h)	641M	130.79 ± 7.96	504.36 ± 57.72	21.912 GB
SAM (ViT-b)	94M	13.59 ± 0.77	133.85 ± 24.26	7.144 GB
SAM-Med2D w/ adapter	271M	1.23 ± 0.07	35.06 ± 12.88	1.489 GB
SAM-Med2D w.o. adapter	91M	0.63 ± 0.02	24.86 ± 9.56	734 MB
MIDeepSeg	3M	0.08 ± 0.02	65.75 ± 21.87	11 MB
ScribblePrompt-UNet	4M	0.27 ± 0.04	1.96 ± 0.20	125 MB

I Ablations

We conduct two ablations of important ScribblePrompt design decisions: (1) synthetic label inputs used during training, and (2) types of prompts simulated during training. We report results on the validation splits of nine validation datasets that were unseen during training.

I.1 Synthetic Labels

Setup. We trained ScribblePrompt-UNet and ScribblePrompt-SAM with different values of p_{synth}, the probability of sampling a synthetic label.

Results. Training with some synthetic labels improves both ScribblePrompt-UNet and ScribblePrompt-SAM's performance on validation data from nine (validation) datasets not seen during training (Fig. 29, 30). For both ScribblePrompt-UNet and ScribblePrompt-SAM, training with 50% synthetic labels leads to the highest Dice on unseen datasets at inference time.

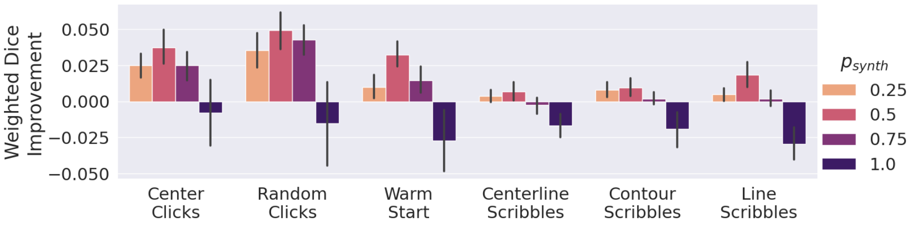


Fig. 29: Probability of synthetic labels during training for ScribblePrompt-UNet. We report change in Dice relative to ScribblePrompt-UNet trained without any synthetic labels ($p_{synth} = 0$). We show Dice after five steps of simulated interactions following six different (inference-time) interaction procedures. Errorbars show 95% CI.

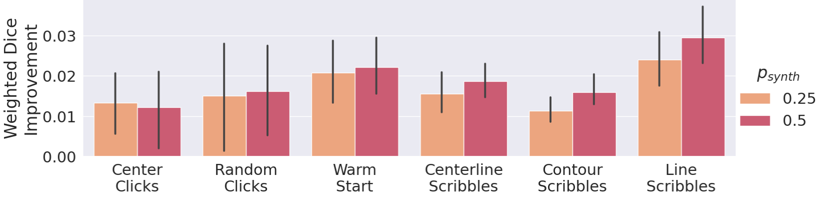


Fig. 30: Probability of synthetic labels during training for ScribblePrompt-SAM. We report change in Dice relative to ScribblePrompt-SAM trained without any synthetic labels ($p_{synth} = 0$). We show Dice after five steps of simulated interactions following six different (inference-time) interaction procedures. Errorbars show 95% CI.

I.2 Prompt Types

Setup. We evaluate ScribblePrompt-UNet models trained with different combinations of prompts, compared to the complete ScribblePrompt-UNet:

- **ScribblePrompt-UNet (scribbles)** trained on boxes and scribbles.
- **ScribblePrompt-UNet (clicks)** trained on boxes and clicks.
- **ScribblePrompt-UNet (random clicks)** trained on boxes and random clicks.

Results. Fig. 31 shows results for six different inference-time interaction procedures. ScribblePrompt-UNet trained with scribbles, clicks, and bounding boxes predicts segmentations more accurately than do ablated versions of ScribblePrompt-UNet.

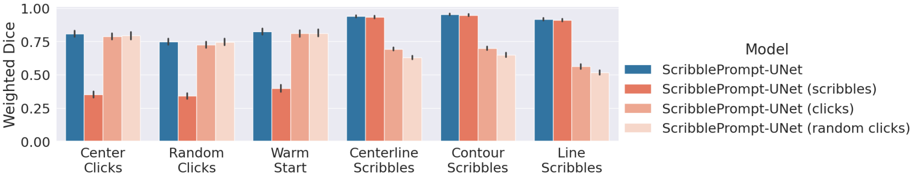


Fig. 31: Ablation of interactions during training. We report Dice after five steps of simulated interactions following six inference-time interaction procedures. Error bars show 95% CI from bootstrapping.

References

1. Abdi, A.H., Kasaei, S., Mehdizadeh, M.: Automatic segmentation of mandible in panoramic x-ray. Journal of Medical Imaging **2**(4), 044003 (2015) 9, 12
2. Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., Zou, J.Y.: Gradio: Hassle-free sharing and testing of ML models in the wild. CoRR **abs/1906.02569** (2019), http://arxiv.org/abs/1906.02569 12, 32
3. Aine, C.J., Bockholt, H.J., Bustillo, J.R., Cañive, J.M., Caprihan, A., Gasparovic, C., Hanlon, F.M., Houck, J.M., Jung, R.E., Lauriello, J., Liu, J., Mayer, A.R., Perrone-Bizzozero, N.I., Posse, S., Stephen, J.M., Turner, J.A., Clark, V.P., Calhoun, V.D.: Multimodal Neuroimaging in Schizophrenia: Description and Dissemination. Neuroinformatics **15**(4), 343–364 (Oct 2017). https://doi.org/10.1007/s12021-017-9338-9, http://link.springer.com/10.1007/s12021-017-9338-9 9, 26
4. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2020). https://doi.org/https://doi.org/10.1016/j.dib.2019.104863, https://www.sciencedirect.com/science/article/pii/S2352340919312181 8, 9, 24
5. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016) 2
6. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021) 10
7. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017) 10
8. Bano, S., Vasconcelos, F., Shepherd, L.M., Vander Poorten, E., Vercauteren, T., Ourselin, S., David, A.L., Deprest, J., Stoyanov, D.: Deep placental vessel segmentation for fetoscopic mosaicking. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 763–773. Springer (2020) 10
9. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., 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) 8, 9, 12, 15, 16
10. Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., et al.: The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056 (2019) 10
11. Bloch, N., Madabhushi, A., Huisman, H., Freymann, J., Kirby, J., Grauer, M., Enquobahrie, A., Jaffe, C., Clarke, L., Farahani, K.: Nci-isbi 2013 challenge: automated segmentation of prostate structures. The Cancer Imaging Archive **370**(6), 5 (2015) 10
12. Buda, M., Saha, A., Mazurowski, M.A.: Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Computers in biology and medicine **109**, 218–225 (2019) 10
13. Butoi*, V.I., Ortiz*, J.J.G., Ma, T., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Universeg: Universal medical image segmentation. In: ICCV (2023) 7, 11

14. Caicedo, J.C., Goodman, A., Karhohs, K.W., Cimini, B.A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., Doan, M., McQuin, C., Rohban, M., Singh, S., Carpenter, A.E.: Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. Nature Methods **16**(12), 1247–1253 (Dec 2019). https://doi.org/10.1038/s41592-019-0612-7, https://doi.org/10.1038/s41592-019-0612-7 10
15. Cardona, A., Saalfeld, S., Preibisch, S., Schmid, B., Cheng, A., Pulokas, J., Tomancak, P., Hartenstein, V.: An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. PLoS biology **8**(10), e1000502 (2010) 10
16. Codella, N.C.F., Gutman, D.A., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N.K., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC). CoRR **abs/1710.05006** (2017), http://arxiv.org/abs/1710.05006 10
17. Dalca, A.V., Guttag, J., Sabuncu, M.R.: Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9290–9299 (2018) 9, 26
18. Dalca, A.V., Guttag, J., Sabuncu, M.R.: Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9290–9299 (2018) 10
19. Decenciere, E., Cazuguel, G., Zhang, X., Thibault, G., Klein, J.C., Meyer, F., Marcotegui, B., Quellec, G., Lamard, M., Danno, R., et al.: Teleophta: Machine learning and image processing methods for teleophthalmology. Irbm **34**(2), 196–203 (2013) 10
20. Degerli, A., Zabihi, M., Kiranyaz, S., Hamid, T., Mazhar, R., Hamila, R., Gabbouj, M.: Early detection of myocardial infarction in low-quality echocardiography. IEEE Access **9**, 34442–34453 (2021) 10
21. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945) 2
22. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International journal of computer vision **59**, 167–181 (2004) 6
23. Fischl, B.: Freesurfer. Neuroimage **62**(2), 774–781 (2012) 9, 26
24. Gamper, J., Koohbanani, N., Benes, K., Graham, S., Jahanifar, M., Khurram, S., Azam, A., Hewitt, K., Rajpoot, N.: Pannuke dataset extension, insights and baselines. arxiv. 2020 doi: 10.48550. ARXIV (2003) 10
25. Gerhard, S., Funke, J., Martel, J., Cardona, A., Fetter, R.: Segmented anisotropic ssTEM dataset of neural tissue (11 2013). https://doi.org/10.6084/m9.figshare.856713.v1, https://figshare.com/articles/dataset/Segmented_anisotropic_ssTEM_dataset_of_neural_tissue/856713 10
26. van Ginneken, B., Stegmann, M.B., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Medical Image Analysis **10**(1), 19–40 (2006). https://doi.org/https://doi.org/10.1016/j.media.2005.02.002, https://www.sciencedirect.com/science/article/pii/S1361841505000368 9, 20
27. Gollub, R.L., Shoemaker, J.M., King, M.D., White, T., Ehrlich, S., Sponheim, S.R., Clark, V.P., Turner, J.A., Mueller, B.A., Magnotta, V., et al.: The mcic

collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. Neuroinformatics **11**, 367–388 (2013) 10

28. Gotkowski, K., Lüth, C., Jäger, P.F., Ziegler, S., Krämer, L., Denner, S., Xiao, S., Disch, N., Maier-Hein, K.H., Isensee, F.: Embarrassingly simple scribble supervision for 3d medical segmentation. arXiv preprint arXiv:2403.12834 (2024) 16
29. Gousias, I.S., Edwards, A.D., Rutherford, M.A., Counsell, S.J., Hajnal, J.V., Rueckert, D., Hammers, A.: Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants. Neuroimage **62**(3), 1499–1509 (2012) 10
30. Gousias, I.S., Rueckert, D., Heckemann, R.A., Dyet, L.E., Boardman, J.P., Edwards, A.D., Hammers, A.: Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest. Neuroimage **40**(2), 672–684 (2008) 10
31. Grøvik, E., Yi, D., Iv, M., Tong, E., Rubin, D., Zaharchuk, G.: Deep learning enables automatic detection and segmentation of brain metastases on multisequence mri. Journal of Magnetic Resonance Imaging **51**(1), 175–182 (2020) 10
32. Gut, D.: X-ray images of the hip joints **1** (Jul 2021). https://doi.org/10.17632/zm6bxzhmfz.1, https://data.mendeley.com/datasets/zm6bxzhmfz/1, publisher: Mendeley Data 9, 12, 27
33. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis p. 101821 (2020) 10
34. Hernandez Petzsche, M.R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L., Kofler, F., Ezhov, I., et al.: Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific data **9**(1), 762 (2022) 10
35. Hoopes, A., Hoffmann, M., Greve, D.N., Fischl, B., Guttag, J., Dalca, A.V.: Learning the effect of registration hyperparameters with hypermorph. Machine Learning for Biomedical Imaging **1**, 1–30 (2022), https://melba-journal.org/2022:003 10, 12
36. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical imaging **19**(3), 203–210 (2000) 10, 12
37. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. pmlr (2015) 2
38. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z, http://www.nature.com/articles/s41592-020-01008-z 8, 21
39. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022) 10
40. Karim, R., Housden, R.J., Balasubramaniam, M., Chen, Z., Perry, D., Uddin, A., Al-Beyatti, Y., Palkhi, E., Acheampong, P., Obom, S., et al.: Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge. Journal of Cardiovascular Magnetic Resonance **15**(1), 1–17 (2013) 10

41. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., Baydar, B., Lachinov, D., Han, S., Pauli, J., Isensee, F., Perkonigg, M., Sathish, R., Rajan, R., Sheet, D., Dovletov, G., Speck, O., Nürnberger, A., Maier-Hein, K.H., Bozdağı Akar, G., Ünal, G., Dicle, O., Selver, M.A.: CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021). https://doi.org/https://doi.org/10.1016/j.media.2020.101950, https://www.sciencedirect.com/science/article/pii/S1361841520303145 10, 12
42. Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S.: CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data (Apr 2019). https://doi.org/10.5281/zenodo.3362844, https://doi.org/10.5281/zenodo.3362844 10, 12
43. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 11
44. Kiranyaz, S., Degerli, A., Hamid, T., Mazhar, R., Ahmed, R.E.F., Abouhasera, R., Zabihi, M., Malik, J., Hamila, R., Gabbouj, M.: Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection. IEEE Access **8**, 210301–210317 (2020) 10
45. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: ICCV (2023) 5, 6, 7, 17, 18
46. Krönke, M., Eilers, C., Dimova, D., Köhler, M., Buschner, G., Schweiger, L., Konstantinidou, L., Makowski, M., Nagarajah, J., Navab, N., et al.: Tracked 3d ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. Plos one **17**(7), e0268550 (2022) 10
47. Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., et al.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE transactions on medical imaging **38**(11), 2556–2568 (2019) 10
48. Kuklisova-Murgasova, M., Aljabar, P., Srinivasan, L., Counsell, S.J., Doria, V., Serag, A., Gousias, I.S., Boardman, J.P., Rutherford, M.A., Edwards, A.D., et al.: A dynamic 4d probabilistic atlas of the developing brain. NeuroImage **54**(4), 2750–2763 (2011) 10
49. Lambert, Z., Petitjean, C., Dubray, B., Kuan, S.: Segthor: segmentation of thoracic organs at risk in ct images. In: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6. IEEE (2020) 10
50. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault Workshop Challenge. vol. 5, p. 12 (2015) 8, 9, 10, 12
51. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE transactions on medical imaging **38**(9), 2198–2210 (2019) 10, 12
52. Lemaître, G., Martí, R., Freixenet, J., Vilanova, J.C., Walker, P.M., Meriaudeau, F.: Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. Computers in biology and medicine **60**, 8–31 (2015) 10

53. Li, M., Zhang, Y., Ji, Z., Xie, K., Yuan, S., Liu, Q., Chen, Q.: Ipn-v2 and octa-500: Methodology and dataset for retinal image segmentation. arXiv preprint arXiv:2012.07261 (2020) 10, 12
54. Li, Z., Zheng, Y., Luo, X., Shan, D., Hong, Q.: Scribblevc: Scribble-supervised medical image segmentation with vision-class embedding. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3384–3393 (2023) 16
55. Li, Z., Zheng, Y., Shan, D., Yang, S., Li, Q., Wang, B., Zhang, Y., Hong, Q., Shen, D.: Scribformer: Transformer makes cnn work better for scribble-based medical image segmentation. IEEE Transactions on Medical Imaging (2024) 16
56. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3159–3167 (2016) 16
57. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 5
58. Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al.: Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. Medical image analysis **18**(2), 359–373 (2014) 10
59. Liu, Z., Heer, J.: The effects of interactive latency on exploratory visual analysis. IEEE transactions on visualization and computer graphics **20**(12), 2122–2131 (2014) 35
60. Ljosa, V., Sokolnicki, K.L., Carpenter, A.E.: Annotated high-throughput microscopy image sets for validation. Nature methods **9**(7), 637–637 (2012) 10
61. Löffler, M.T., Sekuboyina, A., Jacob, A., Grau, A.L., Scharr, A., El Husseini, M., Kallweit, M., Zimmer, C., Baum, T., Kirschke, J.S.: A vertebral segmentation dataset with fracture grading. Radiology: Artificial Intelligence **2**(4), e190138 (2020) 10
62. in the Loop, H.: Teeth segmentation dataset, https://humansintheloop.org/resources/datasets/teeth-segmentation-dataset/ 10
63. Luo, X., Hu, M., Liao, W., Zhai, S., Song, T., Wang, G., Zhang, S.: Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 528–538. Springer (2022) 16
64. Luo, X., Liao, W., Xiao, J., Song, T., Zhang, X., Li, K., Wang, G., Zhang, S.: Word: Revisiting organs segmentation in the whole abdominal region. arXiv preprint arXiv:2111.02403 (2021) 10
65. Luo, X., Wang, G., Song, T., Zhang, J., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S.: MIDeepSeg: Minimally Interactive Segmentation of Unseen Objects from Medical Images Using Deep Learning. Medical Image Analysis **72**, 102102 (2021) 17
66. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 1–9 (2024) 8, 17, 21
67. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al.: Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. Medical Image Analysis **82**, 102616 (2022) 10
68. Ma, Y., Hao, H., Xie, J., Fu, H., Zhang, J., Yang, J., Wang, Z., Liu, J., Zheng, Y., Zhao, Y.: Rose: a retinal oct-angiography vessel segmentation dataset and new model. IEEE Transactions on Medical Imaging **40**(3), 928–939 (2021). https://doi.org/10.1109/TMI.2020.3042802 10

69. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience **19**(9), 1498–1507 (2007) 10, 12
70. Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., et al.: The parkinson progression marker initiative (ppmi). Progress in neurobiology **95**(4), 629–635 (2011) 10
71. Marzola, F., Van Alfen, N., Doorduin, J., Meiburger, K.M.: Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment. Computers in Biology and Medicine **135**, 104623 (Aug 2021). https://doi.org/10.1016/j.compbiomed.2021.104623, https://linkinghub.elsevier.com/retrieve/pii/S0010482521004170 10
72. Mazurowski, M.A., Clark, K., Czarnek, N.M., Shamsesfandabadi, P., Peters, K.B., Saha, A.: Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with the cancer genome atlas data. Journal of neuro-oncology **133**, 27–35 (2017) 10
73. Menze, B., Joskowicz, L., Bakas, S., Jakab, A., Konukoglu, E., Becker, A., Simpson, A., D, R.: Quantification of uncertainties in biomedical image quantification 2021. 4th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2021) (2021). https://doi.org/https://doi.org/10.5281/zenodo.4575204 10
74. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014) 10
75. Montoya, A., Hasnin, kaggle446, shirzad, Cukierski, W., yffud: Ultrasound nerve segmentation (2016), https://kaggle.com/competitions/ultrasound-nerve-segmentation 10
76. Payette, K., de Dumast, P., Kebiri, H., Ezhov, I., Paetzold, J.C., Shit, S., Iqbal, A., Khan, R., Kottke, R., Grehten, P., et al.: An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. Scientific Data **8**(1), 1–14 (2021) 10
77. Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E.: An open access thyroid ultrasound image database. In: Romero, E., Lepore, N. (eds.) 10th international symposium on medical information processing and analysis. vol. 9287, p. 92870W. SPIE / International Society for Optics and Photonics (2015). https://doi.org/10.1117/12.2073532, https://doi.org/10.1117/12.2073532 10
78. Podobnik, G., Strojan, P., Peterlin, P., Ibragimov, B., Vrtovec, T.: HaN-Seg: The head and neck organ-at-risk CT and MR segmentation dataset. Medical Physics **50**(3), 1917–1927 (2023). https://doi.org/https://doi.org/10.1002/mp.16197, https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.16197, tex.eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.16197 10
79. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid) (2018). https://doi.org/10.21227/H25W98, https://dx.doi.org/10.21227/H25W98 10
80. Radau, P., Lu, Y., Connelly, K., Paul, G., Dick, A., Wright, G.: Evaluation framework for algorithms segmenting short axis cardiac mri. The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge **49** (2009) 9, 12, 19

81. Rakic, M., Wong, H.E., Ortiz, J.J.G., Cimini, B., Guttag, J.V., Dalca, A.V.: Tyche: Stochastic in-context learning for medical image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2024) 7, 11
82. Rister, B., Yi, D., Shivakumar, K., Nobashi, T., Rubin, D.L.: CT-ORG, a new dataset for multiple organ segmentation in computed tomography. Scientific Data **7**(1), 381 (Nov 2020). https://doi.org/10.1038/s41597-020-00715-8, https://www.nature.com/articles/s41597-020-00715-8 10
83. Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S., Nguyen, C., Ngo, V.D., Seekins, J., Blankenberg, F.G., Ng, A., et al.: Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. MedRxiv (2021) 10
84. Seibold, C., Reiß, S., Sarfraz, S., Fink, M.A., Mayer, V., Sellner, J., Kim, M.S., Maier-Hein, K.H., Kleesiek, J., Stiefelhagen, R.: Detailed annotations of chest x-rays via ct projection for report understanding. In: Proceedings of the 33th British Machine Vision Conference (BMVC) (2022) 10, 12
85. Serag, A., Aljabar, P., Ball, G., Counsell, S.J., Boardman, J.P., Rutherford, M.A., Edwards, A.D., Hajnal, J.V., Rueckert, D.: Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. Neuroimage **59**(3), 2255–2265 (2012) 10
86. Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis **42**, 1–13 (2017) 10
87. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019) 10
88. Song, Y., Zheng, J., Lei, L., Ni, Z., Zhao, B., Hu, Y.: CT2US: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. Ultrasonics **122**, 106706 (2022). https://doi.org/https://doi.org/10.1016/j.ultras.2022.106706, https://www.sciencedirect.com/science/article/pii/S0041624X22000191 10
89. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE transactions on medical imaging **23**(4), 501–509 (2004) 9, 25
90. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016) 2
91. Vitale, S., Orlando, J.I., Iarussi, E., Larrabide, I.: Improving realism in patient-specific abdominal ultrasound simulation using cyclegans. International journal of computer assisted radiology and surgery **15**(2), 183–192 (2020) 10, 12
92. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023) 7, 8, 9, 28
93. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018) 2
94. Ye, J., Cheng, J., Chen, J., Deng, Z., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. arXiv preprint arXiv:2311.11969 (2023) 8

95. Zhang, K., Zhuang, X.: Cyclemix: A holistic strategy for medical image segmentation from scribble supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11656–11665 (2022) 16
96. Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. Commun. ACM **27**(3), 236–239 (mar 1984). https://doi.org/10.1145/357994.358023, https://doi.org/10.1145/357994.358023 3
97. Zhang, Y., Xian, M., Cheng, H.D., Shareef, B., Ding, J., Xu, F., Huang, K., Zhang, B., Ning, C., Wang, Y.: Busis: A benchmark for breast ultrasound image segmentation. In: Healthcare. vol. 10, p. 729. MDPI (2022) 10
98. Zhao, Q., Lyu, S., Bai, W., Cai, L., Liu, B., Wu, M., Sang, X., Yang, M., Chen, L.: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. CoRR **abs/2207.06799** (2022) 10
99. Zheng, G., Chu, C., Belavỳ, D.L., Ibragimov, B., Korez, R., Vrtovec, T., Hutt, H., Everson, R., Meakin, J., Andrade, I.L., et al.: Evaluation and comparison of 3d intervertebral disc localization and segmentation methods for 3d t2 mr data: A grand challenge. Medical image analysis **35**, 327–344 (2017) 9, 12
100. Zheng, X., Wang, Y., Wang, G., Liu, J.: Fast and robust segmentation of white blood cell images by self-supervised learning. Micron **107**, 55–71 (2018). https://doi.org/https://doi.org/10.1016/j.micron.2018.01.010, https://www.sciencedirect.com/science/article/pii/S0968432817303037 9, 12