Supplementary Information

PropSAM: A Propagation-Based Model for Segmenting Any 3D Objects in Multi-Modal Medical Images

Zifan Chen[1*], Xinyu Nan[1*], Jiazheng Li[2*], Jie Zhao[3], Haifeng Li[4], Zilin Lin[1], Haoshen Li[1], Heyun Chen[1], Yiting Liu[2], Bin Dong[4,5,6✉], Li Zhang[1,4✉], and Lei Tang[2✉]

[1]Center for Data Science, Peking University, Beijing, China

[2]Department of Radiology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing, China

[3]National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing, China

[4]National Biomedical Imaging Center, Peking University, Beijing, China

[5]Beijing International Center for Mathematical Research (BICMR), Peking University, Beijing, China

[6]Center for Machine Learning Research, Peking University, Beijing, China

August 25, 2024

Contents

Supplementary texts

Supplementary figures

Supplementary tables

Supplementary texts

S1 Methodology details of PropSAM

As shown in Supplementary Figure S1, our proposed PropSAM is designed to accept user prompts in the forms of bounding-box-style or 2D-mask-style (Supplementary Figure S1A), facilitating efficient segmentation of any 3D object in any modalities (Supplementary Figure S1B). Initially, users load 3D images from various modalities, such as computed tomography (CT), magnetic resonance imaging (MRI), or positron emission tomography-computed tomography (PET-CT), identify their objects of interest, and provide prompts accordingly. If the prompt is a bounding box, it is first converted into a 2D mask using the Box2Mask module to serve as a 2D-mask-style prompt (Supplementary Figure S1C). This 2D-mask slice then acts as the guiding prompt for the segmentation of adjacent slices via the PropMask module, which leverages information propagation among these slices (Supplementary Figure S1E). Subsequently, the boundary slices from the propagation are utilized as prompt slices in subsequent propagation rounds until no further segmentation results are predicted, resulting in a 3D mask of the target object (Supplementary Figure S1F). This section details the components of this workflow in the following.

S1.1 Two styles of prompts

Our proposed PropSAM is designed to support two styles of prompts:

Style 1: Bounding box (less user interaction) This prompt style enables users to specify a bounding box for the target object on a slice, typically on the transverse plane of medical 3D images, as illustrated in Supplementary Figure S2A. Unlike the recently popular models MedSAM [1] and SegVol [2], which require users to provide bounding boxes on two planes—typically the transverse and sagittal planes, effectively creating a 3D bounding box—our PropSAM requires only a single-plane bounding box, truly a 2D prompt (Supplementary Figure S2B). This difference not only reduces manual interaction but also aligns more closely with typical clinical practices, where transverse planes are commonly used for review.

Style 2: 2D pixel-level mask At the core of PropSAM is the use of a 2D segmented slice of the target object as the initial guiding slice, which facilitates the generation of segmentations for adjacent slices by propagating information between them. For prompts based on the bounding box style, our Box2Mask module is employed to convert these into 2D mask-style prompts, ensuring uniformity in the inputs for subsequent processing. Users can also directly provide a 2D mask segmentation on a slice from any view (Supplementary Figure S2C). While this prompt style requires more user interaction at the pixel level, it provides more precise information in the initial slice, which proves especially beneficial for segmenting challenging samples and objects (refer to the "Results" section in our paper).

S1.2 Inputs of PropSAM

The aim of PropSAM is to segment any 3D medical images effectively, including but not limited to modalities such as CT, MR, and PET-CT. To facilitate this, we have compiled a collection of 43 public 3D medical segmentation datasets across multiple modalities (refer to Supplementary Table S1). Utilizing these datasets, we developed PropSAM, which involved the initial training of a Box2Mask module (described in Supplementary Text S1.3) followed by the training of a PropMask module (detailed in Supplementary Text S1.4). The functionalities and operational details of these two modules are elaborated in subsequent sections.

S1.3 Details of Box2Mask module

As outlined in our PropSAM workflow (Supplemenetary Figure S1), when a user inputs a bounding box prompt, it is initially processed by the Box2Mask module. This module converts the bounding box into a mask prompt, standardizing the format for subsequent input into the PropMask module. In this section, we detail the experimental settings of Box2Mask module, including module architecture, data preprocessing and characteristics, training configurations, and inference settings.

S1.3.1 Architecture of Box2Mask module

The objective of this module is to transform Region of Interest (ROI) images, which are cropped using bounding box prompts, into foreground masks (refer to Supplementary Figure S[TODO]). This process necessitates a focus on local features. Moreover, to enhance the efficiency of this bounding box to mask conversion stage, we have selected a convolutional neural network (CNN) as the foundational framework. And we note that the UNet-based architecture [3], particularly the nnUNet model [4], has become the most widely adopted and effective approach for medical imaging segmentation in recent years. Therefore, we have decided to build our Box2Mask module based on the UNet architecture.

The Box2Mask module comprises a six-stage encoder-decoder UNet architecture. The input utilizes three-dimensional channels, appropriate for a grayscale image replicated three times. The initial stage features 32 channels, which doubles with each subsequent stage, capping at 512. Thus, the channel counts across the six stages are [32, 64, 128, 256, 512, 512]. All convolutional kernels are 3, with a stride of one within each stage and a stride of two in the last layer of each stage to reduce resolution. Each stage includes two convolutional layers. The normalization layer employs instance normalization [5], and the activation function is LeakyReLU. Each encoder layer is linked to a corresponding decoder layer through a skip connection to maintain low-frequency features. All six decoding stages are designed to produce binary foreground segmentation predictions, which is called deep supervision introduced in the following subsection. The total parameters of this Box2Mask module is 20.62M. The source code for this module is available in the Supplementary Materials, where further details can be accessed.

S1.3.2 Data preprocessing and characteristics

In this study, we collected 43 public 3D medical segmentatin datasets (see Supplemenetary Table S1). Based on these datasets, we developed the Box2Mask module, designed to transform a ROI image into its corresponding foreground mask. The data were prepared through several steps (Supplementary Figure S3): 1) simulation of bounding boxes based on 3D masks to obtain ROI images; 2) ROI image normalization; 3) random data augmentation to enhancing the training ROI images. Below, we detail these steps:

Step 1: Generating bounding boxes (Supplementary Figure S3A). Initialy, we examined the 3D images on one plane, if a slice contained any mask annotation with an area exceeding 100 pixels, we generated the tightest bounding box around it. We then randomly adjusted its width and height with a scaling ratio bewteen 1.0 to 1.25 to account for potential deviation in actual usage, producing what we refer to as an ROI image $\mathbf{x}$.

Step 2: Normalizing ROI images (Supplementary Figure S3B). After acquiring the ROI images, normalizaiton was applied to enhance clarity and emphasize the foreground region of interest. We determined the 0.5th and 99.5th percentiles of pixel values within the annotated mask of the original slice images as the minimum and maximum values, v_{min} and v_{max}, respectively. Each ROI image was then normalized to $\tilde{\mathbf{x}} = 2.0 \times (f(\mathbf{x}) - v_{min})/(v_{max} - v_{min}) - 1.0$, where $f(\mathbf{x}) = \min(v_{max}, \max(v_{min}, \mathbf{x}))$ acts as a clipping function. This process yielded a normalized ROI image and its corresponding annotated mask as a basic training and evaluating sample $(\tilde{\mathbf{x}}, \mathbf{y})$, where $\mathbf{y} \in \{0, 1\}$ denotes the foreground object.

Step 3: Data augmentation (Supplementary Figure S3C). To optimize training efficiently, we applied offline data augmentation five times for each sample. Specially, each image had a 50% chance of being flipped horizontally or vertically. Additionally, we randomly adjusted the image's brightness and contrast, also with a 50% probability, setting the adjustment ranges to [-0.2, 0.2]. The images were also rotated randomly up to 45 degrees with a 50% probability, filling any areas outside the original boundaries with a constant value (typically black). These samples were uniformly resized to a resolution of 224×224 for input into the Box2Mask module.

Following these preprocessing steps, we obtained a total of 19,344,368 samples across 44 datasets comprising 284 objects (see Supplementary Figure S4). According to the data partitioning in MedSAM [1], these data were divided into interval and exterval validation datasets. The interval validation dataset was further split into training and validation sets at an 80:20 ratio, resulting into 14,974,620 training samples, 3,782,206 interval validation samples, and 587,542 exterval validation samples. Supplementary Table S2 presents the detailed characteristics of the data used for the Box2Mask module across these datasets. We trained the Box2Mask module on the training set, with the interval validation set used to evaluate model performance and select the final model checkpoint. The external validation dataset served to demonstrate the robustness of Box2Mask and its zero-shot capability with unseen objects and datasets.

S1.3.3 Training configurations

In this study, we utilized PyTorch [6] (version 2.0.0) to implement our models and executed them on a server equipped with the CUDA platform (version 11.8). The Box2Mask module leverages deep supervision, enabling predictions at six distinct stages, denoted as $\{\mathbf{P}_1,\cdots,\mathbf{P}_S\}_{S=6}$. Each prediction, $\mathbf{P}_s$, activated by the Sigmoid function, outputs a 2D representation where values between $[0.0,1.0]$ indicate the probability that each pixel is part of the foreground. The foreground ground truth is accordingly rescaled to align with the resolutions of these six stages, represented as $\{\mathbf{M}_1,\cdots,\mathbf{M}_S\}_{S=6}$, where each $\mathbf{M}_s$ is binary with 1 indicating the foreground. To calculate the loss, we apply soft dice loss at each stage and then compute the average of these losses to derive the overall loss function, which is expressed as:

$$L_{\text{Box2Mask}} = \frac{1}{S}\sum_{s=1}^{S}\left(1.0 - \frac{2\times\sum_{i=1}^{W_s}\sum_{j=1}^{H_s}\mathbf{P}_{s,i,j}\mathbf{M}_{s,i,j}}{\sum_{i=1}^{W_s}\sum_{j=1}^{H_s}\mathbf{P}_{i,j}^2+\sum_{i=1}^{W_s}\sum_{j=1}^{H_s}\mathbf{M}_{i,j}^2}\right), \tag{1}$$

where W_s and H_s denote the resolution at the sth stage. The loss ranges from 0.0 to 1.0, ensuring that the module is trained effectively across all resolutions, thus enhancing its predictive accuracy and reliability.

The Box2Mask module was trained using four NVIDIA A800-SXM4-80GB GPUs and 64 Intel(R) Xeon(R) Platinum 8358P CPUs (2.60GHz). The AdamW optimizer was utilized with an initial learning rate of 1e-3 as well as a weight decay of 1e-4. The learning rate was adjusted according to a Cosine Annealing LR schedule with a maximum period of 100 epochs and a minimum eta of 1e-5. During each epoch, we randomly selected 10,000 samples for training and conducted evaluations every 20 epochs using a set of 5,000 randomly sampled validation samples. The training lasted for 4,100 epochs, with a batch size of 1,024, over a span of about six days. Supplementary Figure S5 illustrates the training and validation loss curves. We selected the latest checkpoint as the final weight configuration for our Box2Mask module.

S1.3.4 Inference settings

Once we trained the Box2Mask module, we salloc one GPU with 8 CPUs for inference evaluation, as well as the compared methods, to ensure that inference time and resource comparisons fair. For inference phase, we first cropped the ROI images from the promptable bounding boxes, then normalized them by a series of candidated minimum and maximum parameters. The minimum parameters are determined using the 5th to 40th percentiles (in steps of 1), while the maximum parameters are determined using the 90th to 95th percentiles (in steps of 0.5). These parameters are then combined to standardize the ROI images, resulting in candidate normalized ROI images. Subsequently, the Box2Mask module is employed to predict the foreground. The final standardization parameters, v_{min} and v_{max}, are determined based on the 0.5th and 99.5th percentile values of the pixel locations predicted as foreground. These parameters are then used to standardize and predict the ROI image.

S1.4 Details of PropMask module

As outlined in the Supplementary Text S1.3, two styles of prompts are consistently transformed into the 2D mask on a slice of the target object. Subsequently, this 2D mask serves as the initial guiding slice, and the PropMask module utilizes it to generate segmentation for the adjacent slices by leveraging propagation information (refer to Supplementary Figure S1). In this section, we provide a detailed description of the experimental settings for the PropMask module, including its architecture, data preprocessing and characteristics, training configurations, and inference settings.

S1.4.1 Architecture of PropMask module

PropMask serves as the core component of the network and its architecture is largely based on UNet. PropMask consists of an image encoder, a mask encoder, a sequence of cross-attention modules and a decoder.

The image encoder and mask encoder are also six-stage CNN encoders, which are the same as the encoder of Box2Mask module, but the input channel of the mask encoder is one to accept the 2D mask prompt directly. Both the guiding slice and its adjacent slices go through the image encoder to produce support features and query features of six resolutions([224×224, 112×112, 56×56, 28×28, 14×14, 7×7]), respectively.

Similarly, the 2D mask from the guiding slice go through the mask encoder to produce mask features of six resolutions. Subsequently, a sequence of cross-attention modules are employed. Given a set of query vectors Q, key vectors K, and value vectors V, the definition of cross-attention is as follows:

$$\text{Attention}(Q,K,V) = \text{SoftMax}(\frac{QK^T}{\sqrt{d_k}}) \tag{2}$$

where QK^T represents the dot product between queries and keys, which measures the similarity or alignment between the queries and keys and cross-attention is particularly used when the sets of queries, keys and values are derived from different input sources, enabling the model to integrate information across these sources.

Support features, query features, and mask features of PropMask are respectively flattened into 1-dimensional vectors, serving as the support, query and value vectors for cross-attention. Considering the definition of cross-attention, the outputs of the cross-attention modules of PropMask can be regarded as the value vectors for the query features. The output value vectors can be reshaped to 2D feature maps, which serve as the feature maps for query images. To balance the model efficiency and performance, cross-attention modules are executed only on the lowest four resolution feature maps of the slices and mask([56×56, 28×28, 14×14, 7×7]). Finally, the output of cross-attention at the lowest resolution go through multiple de-convolutions in the decoder to generate outputs of varying resolutions, which match the image features of the six-stage encoder. Following the skip connection structure of the UNet architecture, the output from each stage of the decoder is concatenated with the feature maps of the same resolution from either the cross-attention module or the encoder stage to be the input of the next decoder stage and the final output of the decoder is the prediction mask.

S1.4.2 Data preprocessing and characteristics

We utilized the 43 public 3D medical segmentation datasets referenced in Supplementary Text S1.3.2 to train and evaluate our PropMask module. The primary function of the PropMask module is to propagate the 2D mask from a guiding slice to its adjacent slices. Therefore, we defined a task structure for the PropMask module, consisting of a guiding slice $\mathbf{G} = \mathbf{G}_{img}, \mathbf{G}_{mask}$ that includes the original slice image and its corresponding 2D mask. The adjacent slices, $\mathbf{A} = \mathbf{A}_1, \cdots, \mathbf{A}_N$, comprise images around the guiding slice within a 20mm range, with the maximum number N of adjacent slices set at 20. Supplementary Figure S6 illustrates several sample tasks for training and evaluating the PropMask module. These tasks undergo several preparation steps (Supplementary Figure S7) to serve as inputs for the PropMask module: 1) calculating the cropped size to crop both the guiding and adjacent slices, thereby constructing the ROI tasks; 2) normalizing ROI tasks; 3) applying random data augmentation to enhance the training of ROI tasks. Below, we detail these steps:

Step 1: Cropping ROI tasks (Supplemenetary Figure S7A) Initially, we generated the tightest bounding box around the mask of the guiding slice and then randomly adjusted its width and height with a scaling ratio between 1.0 to 2.0 to capture the context around the target object. This adjusted bounding box was then used to crop both the guiding and adjacent slices, forming the cropped ROI tasks.

Step 2: Normalizing ROI tasks (Supplemenetary Figure S7B) Following a similar normalization process as in the Box2Mask module, we determined the 0.5th and 99.5th percentiles of pixel values within the annotated mask of the cropped guiding slice image as the minimum and maximum values, respectively. These parameters were then utilized to normalize both the guiding slice image and adjacent slice images.

Step 3: Data augmentation (Supplemenetary Figure S7C) We implemented online data augmentation during the training of the PropMask module. Specifically, each image in a task had a 50% chance of being flipped horizontally or vertically. Additionally, images were randomly rotated up to 45 degrees with a 50% probability and uniformly resized to a resolution of 224×224 for input into the PropMask module.

Following these preprocessing steps, we obtained a total of 1,345,871 tasks across the 44 datasets. According to the data partitioning in MedSAM, and consistent to the division in the Box2Mask module, these data were segmented into interval and external validation datasets. The interval validation dataset was further divided into training and validation sets at an 80:20 ratio, resulting in 1,020,576 training tasks, 258,889 interval validation tasks, and 66,406 external validation tasks. Supplementary Table S3 provides detailed characteristics of the data used for the PropMask module across these datasets. Please note that the fundamental training unit of PropMask is a task, each containing several images (typically 20 adjacent images and one guding image), and we perform online data augmentation. This is different from the Box2Mask, which, as described in Supplemenetary Text S1.3.2, employs five times offline data augmentation and counts the image as the basic training unit, this difference would result in offline statistical quantitative difference. The PropMask module was trained on the training set, with the interval validation set used to evaluate model performance and select the final model checkpoint. The external validation dataset served to demonstrate the robustness of PropMask and its zero-shot capability with unseen objects and datasets.

S1.4.3 Training configurations

We implemented the PropMask module using PyTorch (version 2.0.0) and executed it on a server equipped with CUDA version 11.8. The PropMask module also utilizes deep supervision, and its overall loss function is similar to that described in Supplementary Text S1.3.3, but extended to accommodate the number of adjacent slices.

The PropMask module was trained on four NVIDIA A800-SXM4-80GB GPUs and 64 Intel(R) Xeon(R) Platinum 8358P CPUs (2.60GHz). We employed the AdamW optimizer, initiating the process with a learning rate of 5e-4 and a weight decay of 1e-4. The learning rate was modulated following a Cosine Annealing LR schedule with a maximum period of 100 epochs and a minimum eta of 1e-5. Throughout the training process, we randomly selected 10,000 tasks per epoch, each task consisting of the guiding slice and four randomly sampled adjacent slices. Evaluations were conducted every 20 epochs using a set of 5,000 randomly selected validation tasks. The training extended over 4,500 epochs, with a batch size of 160, lasting approximately seven days. Supplementary Figure S8 displays the training and validation loss curves. We chose the most recent checkpoint as the final weight configuration for our PropMask module.

S1.4.4 Inference settings

Once the PropMask module was trained, the model acquired the capability to segment slices based on a guiding slice. For this purpose, we allocated one GPU along with 8 CPUs, mirroring the settings used during the inference phase of the Box2Mask module, to conduct inference experiments and ensure fair comparison of inference performance and resource utilization. As illustrated in Supplementary Figure XXX, during the inference phase, we first identified the guiding slice in a 3D medical image. We then considered this slice and its adjacent slices within a 20mm range as the basic task for the input of the PropMask module. The model was tasked with generating the segmentation for these adjacent slices. Subsequently, we determined the boundaries of these adjacent slices to serve as new guiding slices in subsequent rounds, and initiated new tasks for the next PropMask generation. This propagating interaction continued until the model no longer predicted any foreground region. Therefore, we could obtain the 3D segmentatin of the target object indicated in the initial guiding slice.

S2 Details of testing subset sampling

To evaluate the reliability of PropSAMs in practical clinical scenarios, we constructed a testing subset by randomly sampling samples from the dataset and provided it to the doctor to test the interaction efficiency of different interactive segmentation models.

We adhered to the following protocols to collect the testing samples: firstly, we sampled 50 3D medical images in total. Moreover, these samples should contain at least 50 different segmentation objects and each sample should provide at least one segmentation object in the 50 different segmentation objects. Finally, each sample is designated a different segmentation object that is different from other samples.

During the testing experimental analysis, the doctor only need to provide the prompt for the specific segmentation object that we designated for each sample. We recorded the time it took for doctors to complete interaction tests with different prompts from different segmentation models.

287 ## Supplementary figures

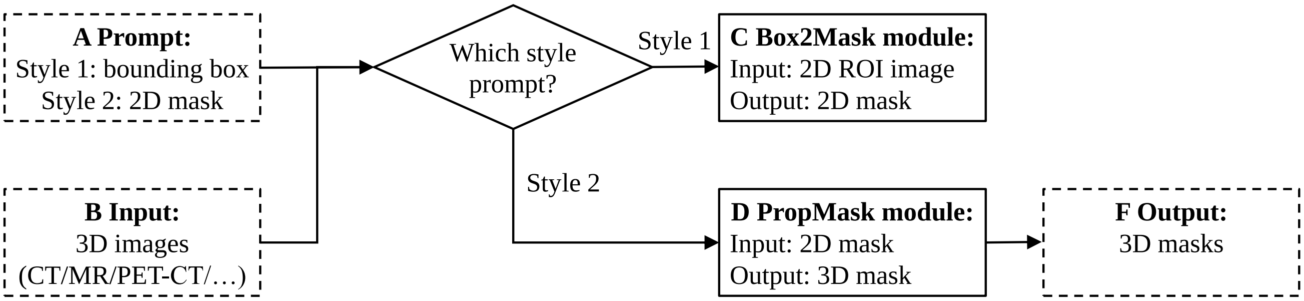

Figure S1: Workflow of the proposed PropSAM.

A Two promptable styles supported by PropSAM. Style 1 is the bounding box, which requires users to specify a bounding box on a view of slices encompassing the target object; Style 2 is the 2D mask, which involves user-segmented, pixel-level delineation of the target object on a slice.

B The input of PropSAM can include any medical 3D iamges, such as CT, MR, PET-CT, and micro-CT.

C The Box2Mask moudle within PropSAM is designed to transform 2D ROI images, dervied from bounding box prompts, into 2D mask-style prompts.

D The PropMask module within PropSAM utilizes 2D mask prompts as the initial guiding slice to segment adjacent slices. This is achieved by propagating information between slices, resulting in a 3D segmentation of the target object.

F The output from PropSAM is a 3D mask of the target object.

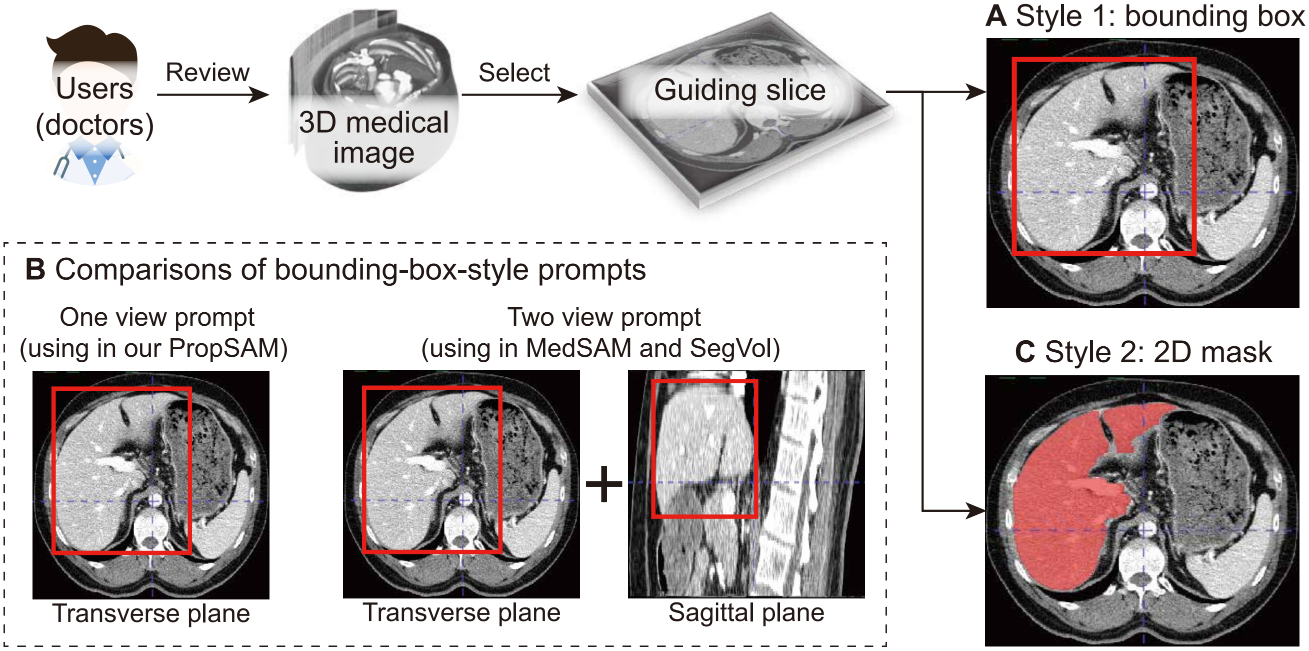

Figure S2: Two styles of prompts in the PropSAM.

A Style 1: users provide a bounding box encompassing the target object on a view of slices in 3D medical images.

B Comparisons between the bounding-box-style prompt of PropSAM and two recently popular models: MedSAM and SegVol.

C Style 2: users provide a 2D pixel-level mask delineation of the target object on a slice.

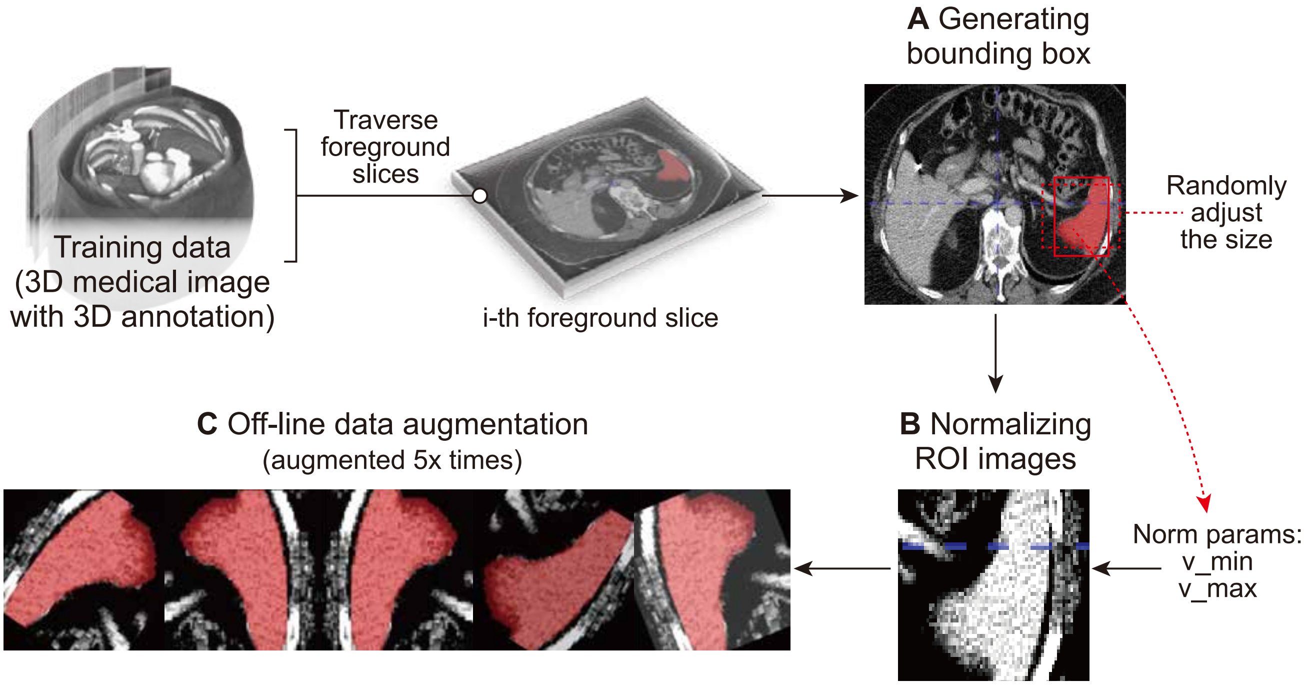
A Generating bounding box
Traverse foreground slices
Training data (3D medical image with 3D annotation)
i-th foreground slice
Randomly adjust the size
C Off-line data augmentation (augmented 5x times)
B Normalizing ROI images
Norm params: v_min v_max

Figure S3: Data preprocessing for Box2Mask module.

A Generating bounding boxes according to 3D annotations.

B Normalizing ROI images by computed normalization parameters based on foreground annotation.

C Offline data augmentation to optimize training efficiently.

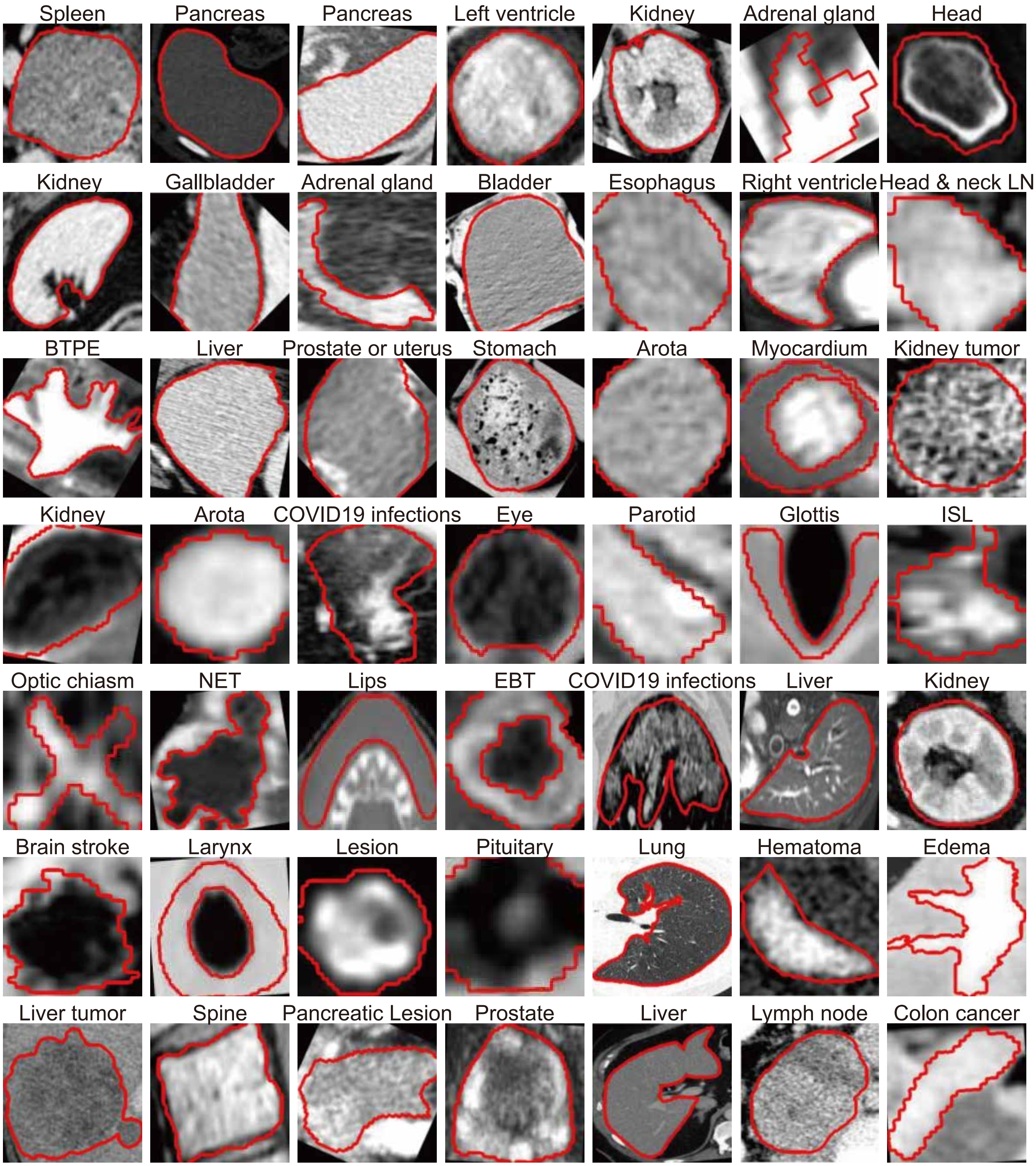

Figure S4: Visualizaiton of input samples of Box2Mask module.

We randomly visualized several samples, each sample contains a ROI image and its corresponding mask (represented by a red boundary in the figure).

*Abbreviation: nonenhancing brain tumor (NBT); enhancing brain tumor (EBT); brain tumor peritumoral edema (BTPE); ischemic stroke lesion (ISL); lymph node (LN).

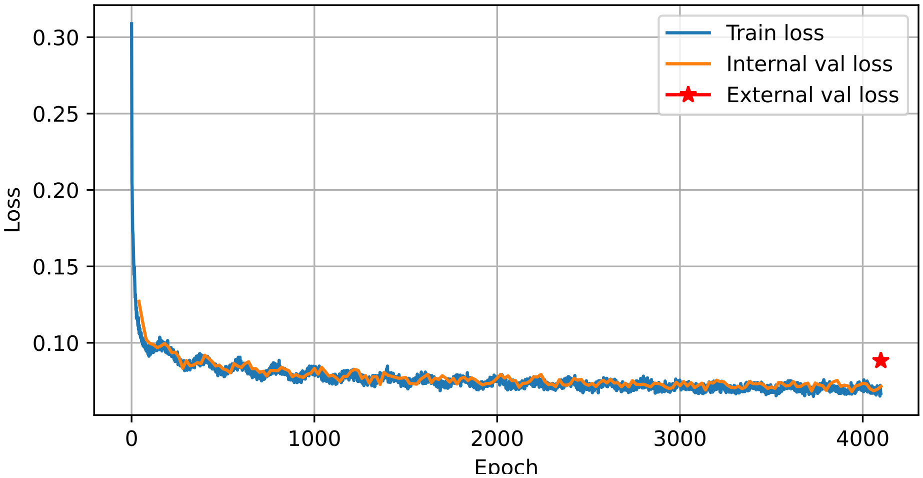

Figure S5: Loss curves of the Box2Mask module.

The blue line depicts the training loss curve of the Box2Mask module across a total of 4,100 epochs, with each epoch involving the training of 10,000 samples. The red line illustrates the interval validation loss, evaluated every 20 epochs using a set of 5,000 randomly sampled validation samples. The red star marks the loss value evaluated by the latest checkpoint on the external validation set.

Figure S6: Visualization tasks for training and evaluating the PropMask module.

The first column represents the guiding slice contains a slice image and its 2D mask. And other columns after that are the adjacent slice images.

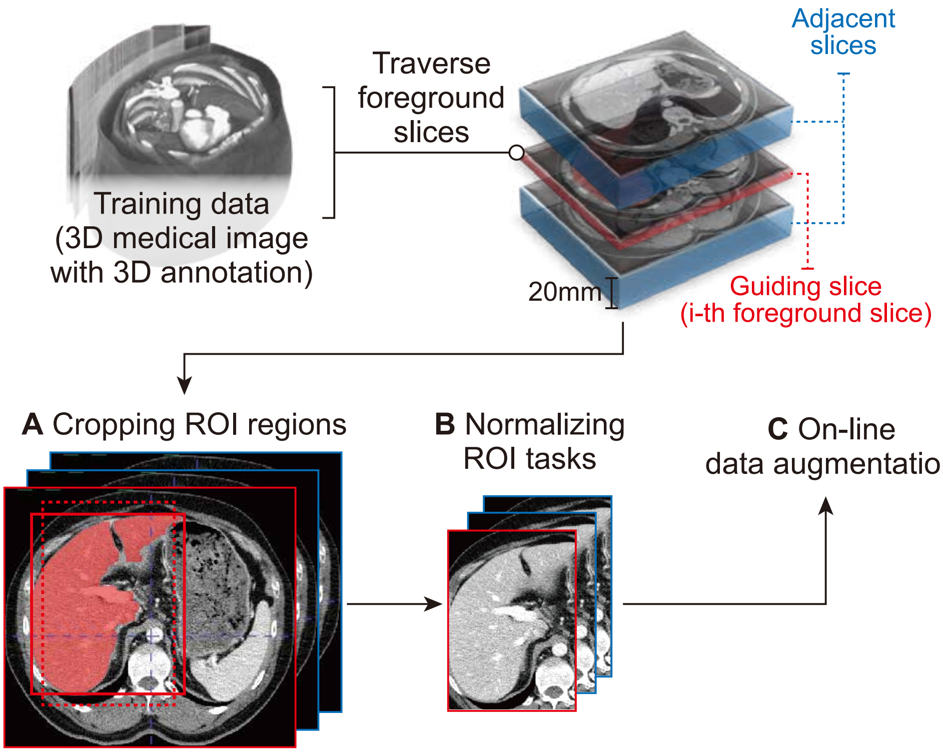
Adjacent slices
Traverse foreground slices
Training data (3D medical image with 3D annotation)
20mm
Guiding slice (i-th foreground slice)
A Cropping ROI regions
B Normalizing ROI tasks
C On-line data augmentation

Figure S7: Data preprocessing for PropMask module.

A Cropping ROI regions from both guiding slice and adjacent slices to construct ROI tasks.

B Normalizing ROI tasks by computed normalization parameters based on foreground region within guiding slice.

C Data augmentation to improve the model's robustness.

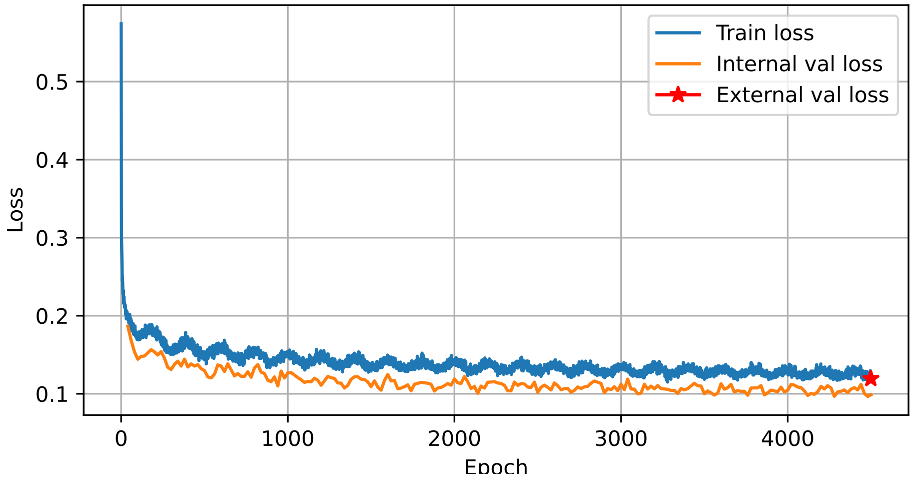

Figure S8: Loss curves of the PropMask module.

The blue line depicts the training loss curve of the PropMask module across a total of 4,500 epochs, with each epoch involving the training of 10,000 tasks. The red line illustrates the interval validation loss, evaluated every 20 epochs using a set of 5,000 randomly sampled validation tasks. The red star marks the loss value evaluated by the latest checkpoint on the external validation set.

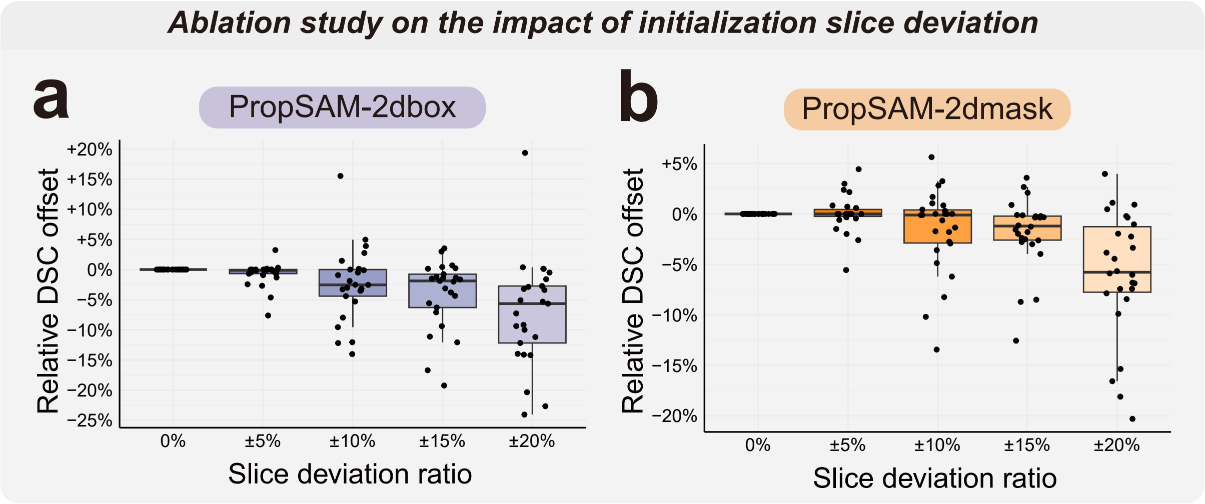

Figure S9: Supplementary results on the impact of initialization slice deviation in PropSAM.

This figure illustrates the effects of initialization slice deviation in PropSAM, using both 2D box and 2D mask prompts, on the percentage fluctuation of relative performance. The x-axis represents the deviation from the maximum slice selected based on RECIST criteria, with deviations of ±5%, ±10%, ±15%, and ±20%. The y-axis shows the performance fluctuation percentage relative to the prompt at the maximum slice (9%).

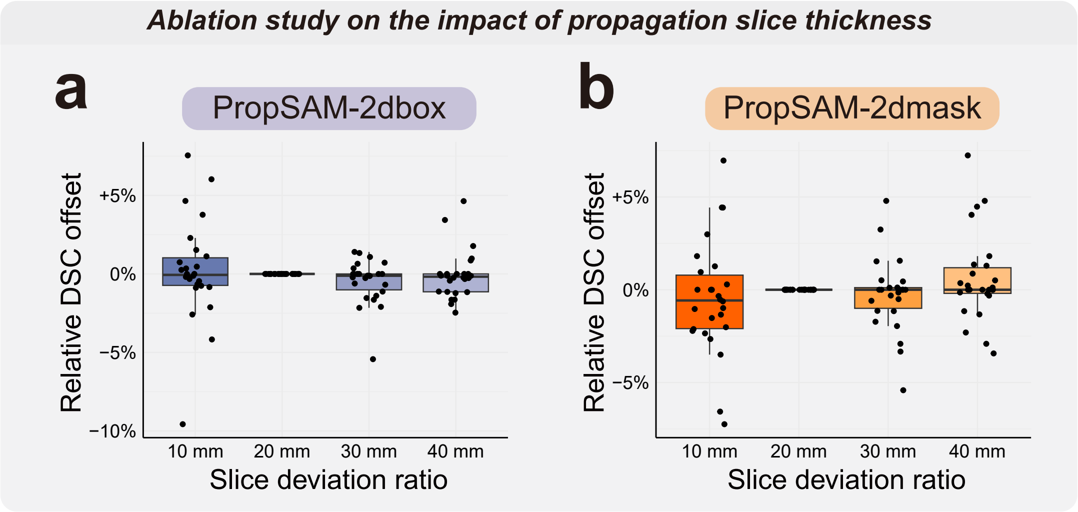

Figure S10: Supplementary results on the impact of propagation slice thickness in PropSAM.

This figure illustrates the effects of propagation slice thickness in PropSAM, using both 2D box and 2D mask prompts, on the percentage fluctuation of relative performance. The x-axis represents the propagation thickness of 10 mm, 20 mm, 30 mm, and 40 mm. The y-axis shows the performance fluctuation percentage relative to the basic thickness (20 mm) empirical selection.

Supplementary tables

Table S1: The dataset used in this study.

Datasets marked with $*$ denote validation datasets and the remaining datasets are internal validation datasets.

ID	Dataset	Modality	Objects	Download link	All	Train.	Val.	Test.
D1	AbdomenCT-1K [7, 8]	CT	Kidneys, liver, pancreas, spleen	https://github.com/JunMa11/AbdomenCT-1K	1	1	1	1
D2	*Adrenal-ACC-Ki67-Seg [9]	CT	Adrenocortical carcinoma	https://doi.org/10.7937/1-VM46	1	1	1	1
D3	AMOS-CT [10]	CT	Arota, bladder, duodenum, esophagus, gallbladder, left kidney, liver, left adrenal gland, prostate or uterus, pancreas, postcava, right kidney, right adrenal gland, spleen, stomach	https://amos22.grand-challenge.org/	1	1	1	1
D4	AutoPET-PETCT [11]	PET-CT	Lesion	https://covid-segmentation.grand-challenge.org/Data/	1	1	1	1
D5	AutoPET-CT [11]	CT	Lesion	https://covid-segmentation.grand-challenge.org/Data/	1	1	1	1
D6	*CHAOS-CT [12]	CT	Liver	https://crossmoda-challenge.ml/	1	1	1	1
D7	COVID-19 Seg. Challenge [13]	CT	COVID-19 infections	https://covid-segmentation.grand-challenge.org/Data/	1	1	1	1
D8	COVID-19-CT-Seg [14]	CT	COVID-19 infections, left lung, right lung	https://github.com/JunMa11/COVID-19-CT-Seg-Benchmark	1	1	1	1

Table S1: (Continued, part 2) The dataset used in this study.

ID	Dataset	Modality	Objects	Download link	All	Train.	Val.	Test.
D9	*HaN-Seg [15]	CT	Arytenoid, brain stem, bone mandible, buccal mucosa, cochleal, cricopharyngeus, cochlear, cavityoral, eyepl, eyepr, esophaguss, eyear, eyeal, glndthyroid, glottis, glndlacrimalr, glndsubmandl, glndlacrimall, glndsubmandr, larynxsg, left acarotid, lips, opticchiasm, opticnrvr, opticnrvl, parotidl, pituitary, parotidr, right acarotid, spinalcord	https://zenodo.org/record/	1	1	1	1
D10	*HCC-TACE-Seg [16, 17]	CT	Liver blood vessel, liver, liver tumor	https://doi.org/10.7937/TCIA.5FNA-0924	1	1	1	1
D11	HECKTOR [18]	PET-CT	Head & neck lymph nodes, head & neck primary tumor	https://hecktor.grand-challenge.org/Overview/	1	1	1	1
D12	INSTANCE [19]	CT	Hematoma	https://instance.grand-challenge.org/	1	1	1	1
D13	KiPA [20, 21]	CT	Kidney tumor, kidney, renal artery, renal vein	https://kipa22.grand-challenge.org/	1	1	1	1
D14	KiTS [22]	CT	Kidney cyst, kidney tumor, kidney	https://kits-challenge.org/kits23/	1	1	1	1
D15	*LNQ2023 [23]	CT	Mediastinal lymph node	https://lnq2023.grand-challenge.org/lnq2023/	1	1	1	1
D16	Lymph nodes [24]	CT	Mediastinal lymph node	https://doi.org/10.7937/K9/TCIA.2015.AQIIDCNM	1	1	1	1

Table S1: (Continued, part 3) The dataset used in this study.

ID	Dataset	Modality	Objects	Download link	All	Train.	Val.	Test.
D17	NSCLC Pleural Effusion [16, 25, 26]	CT	Effusions, thoracic cavitics	https://doi.org/10.7937/tcia.2020.6c7y-gq39	1	1	1	1
D18	*QUBIQ [27]	CT	Pancreatic lesion, pancreas	https://qubiq21.grand-challenge.org/	1	1	1	1
D19	MSD-Task03 Liver [28]	CT	Liver, liver cancer	http://medicaldecathlon.com/	1	1	1	1
D20	MSD-Task06 Lung [28]	CT	Lung cancer	http://medicaldecathlon.com/	1	1	1	1
D21	MSD-Task07 Pancreas [28]	CT	Pancreas cancer, pancreas	http://medicaldecathlon.com/	1	1	1	1
D22	MSD-Task08 HepaticVessel [28]	CT	Hepatic tumour, hepatic vessel	http://medicaldecathlon.com/	1	1	1	1
D23	MSD-Task09 Spleen [28]	CT	Spleen	http://medicaldecathlon.com/	1	1	1	1
D24	MSD-Task10 Colon [28]	CT	Colon cancer primaries	http://medicaldecathlon.com/	1	1	1	1

Table S1: (Continued, part 4) The dataset used in this study.

ID	Dataset	Modality	Objects	Download link	All	Train.	Val.	Test.
D25	Total Segmentator [29]	CT	Adrenalglandleft, adrenalglandright, aorta, autochthonleft, autochthonright, brain, claviculaleft, claviculаright, colon, duodenum, esophagus, face, femurleft, femurright, gallbladder, gluteusmaximusleft, gluteusmaximusright, gluteusmediusleft, gluteusmediusright, gluteusminimusleft, gluteusminimusright, heartatriumleft, heartatriumright, heartmyocardium, heartventricleleft, heartventricleright, hipleft, hipright, humerusleft, humerusright, iliacarteryleft, iliacarteryright, iliacvenaleft, iliacvenaright, iliopsoasleft, iliopsoasright, inferiorvenacava, kidneyleft, kidneyright, liver, lunglowerlobeleft, lunglowerloberight, lungmiddleloberight, lungupperlobeleft, lungupperloberight, pancreas, portalveinandsplenicvein, pulmonaryartery, ribleft, ribright, sacrum, scapulaleft, scapularight, smallbowel, spleen, stomach, trachea, urinarybladder, vertebraec, vertebrael, vertebraet	https://zenodo.org/record/6802614	1	1	1	1
D26	*WORD [30]	CT	Adrenal, bladder, colon, duodenum, esophagus, gallbladder, head, intestine, liver, left kidney, pancreas, right kidney, rectum, spleen, stomach	https://github.com/HiLab-git/WORD	1	1	1	1

Table S1: (Continued, part 5) The dataset used in this study.

ID	Dataset	Modality	Objects	Download link	All	Train.	Val.	Test.
D27	*ACDC [31]	MR	Left ventricle, myocardium, right ventricle	https://humanheart-project.creatis.insa-lyon.fr/database/	1	1	1	1
D28	AMOS-MR [10]	MR	Arota, bladder, duodenum, esophagus, gallbladder, left kidney, liver, left adrenal gland, prostate or uterus, pancreas, postcava, right kidney, right adrenal gland, slpeen, stomach	https://amos22.grand-challenge.org/	1	1	1	1
D29	ATLAS-R2.0 [32]	MR-T1	Brain stroke	https://atlas.grand-challenge.org/	1	1	1	1
D30	BraTS [33]	MR-T1, MR-T1CE, MR-T2, MR-FLAIR	Enhancing brain tumor, brain tumor pertiumoral edema, noenhancing brain tumor core	http://braintumorsegmentation.org/	1	1	1	1
D31	*CHAOS-MR [12]	MR-T1, MR-T2	Left kidney, liver, right kidney, spleen	https://chaos.grand-challenge.org/	1	1	1	1
D32	ISLES [34]	MR-DWI, MR-ADC, MR-FLAIR	Ischemic stroke lesion	http://www.isles-challenge.org/	1	1	1	1
D33	MnM2 [35]	MR	Left ventricle, myocardium, right ventricle	https://www.ub.edu/mnms-2/	1	1	1	1
D34	NCI-ISBI [36]	MR-ADC, MR-T2	Prostate central gland, prostate peripheral	http://dx.doi.org/10.7937/K9/TCIA.2015.zF0vlOPv	1	1	1	1
D35	PI-CAI [37]	MR-bp	Prostate cancer	http://github.com/DIAGNijmegen/picai_labels	1	1	1	1

Table S1: (Continued, part 6) The dataset used in this study.

ID	Dataset	Modality	Objects	Download link	All	Train.	Val.	Test.
D35	PI-CAI [37]	MR-bp	Prostate cancer	http://github.com/DIAGNijmegen/picai_labels	1	1	1	1
D36	PROMISE [38]	MR-T2	Prostate	https://promise12.grand-challenge.org/Details/	1	1	1	1
D37	Qin-Prostate-Repeatability [39, 40]	MR	Prostate gland peripheral zone, prostate suspected tumor, prostate gland, prostate	http://doi.org/10.7937/K9/TCIA.2018.MR1CKGND	1	1	1	1
D38	Spine [41]	MR	Sacral spine, lumbar spine, thoracic spine	https://www.cg.informatik.uni-siegen.de/en/spine-segmentation-and-analysis	1	1	1	1
D39	MSD-Task01 BrainTumour [28]	MR	Edema, enhancing tumor, nonenhancing tumor	http://medicaldecathlon.com/	1	1	1	1
D40	MSD-Task02 Heart [28]	MR	Left atrium	http://medicaldecathlon.com/	1	1	1	1
D41	MSD-Task04 Hippocampus [28]	MR	Anterior, posterior	http://medicaldecathlon.com/	1	1	1	1
D42	MSD-Task05 Prostsate [28]	MR	Peripheral zone, transitional zone	http://medicaldecathlon.com/	1	1	1	1
D43	WMH [42]	MR-T1, MR-FLAIR	Other pathology, white matter hyperintensities	https://wmh.isi.uu.nl/	1	1	1	1

Table S2: Detailed data characteristics of the Box2Mask module across 44 datasets.

The table includes segmentation objects contained in each dataset, the number of 2D images used, and the distribution of these images across training (Train.), validation (Val.), and testing (Test.) sets.

Dataset	Object	All	Train.	Val.	Test.
AbdomenCT-1K-CT		1,492,171	1,189,828	302,343	
	Kidneys	535,052	427,839	107,213	
	Liver	460,650	367,047	93,603	
	Pancreas	261,228	206,830	54,398	
	Spleen	235,241	188,112	47,129	
Adrenal-ACC-Ki67-Seg-CT		8,749			8,749
	Adrenocortical carcinoma	8,749			8,749
AMOS-CT		621,783	516,326	105,457	
	Arota	110,137	91,261	18,876	
	Bladder	20,273	16,768	3,505	
	Duodenum	38,279	31,850	6,429	
	Esophagus	42,481	34,930	7,551	
	Gallbladder	16,573	13,723	2,850	
	Left kidney	42,002	34,904	7,098	
	Liver	61,383	51,050	10,333	
	Left adrenal gland	12,907	10,895	2,012	
	Prostate or uterus	18,205	15,225	2,980	
	Pancreas	32,380	26,944	5,436	
	Postcava	94,357	78,401	15,956	
	Right kidney	40,883	33,909	6,974	
	Right adrenal gland	10,439	8,755	1,684	
	Spleen	38,410	31,687	6,723	
	Stomach	43,074	36,024	7,050	
AutoPET-PETCT		16,481	12,687	3,794	
	Lesion	16,481	12,687	3,794	
AutoPET-CT		16,877	13,083	3,794	
	Lesion	16,877	13,083	3,794	
CHAOS-CT		11,646			11,646
	Liver	11,646			11,646
COVID-19 Seg. Challenge-CT		38,288	30,744	7,544	
	COVID-19 infections	38,288	30,744	7,544	
COVID-19-CT-Seg-CT		42,337	36,723	5,614	
	COVID-19 infections	19,143	17,678	1,465	
	Left lung	12,453	10,115	2,338	
	Right lung	10,741	8,930	1,811	

Table S2: (Continued, part 2) Detailed data characteristics of the Box2Mask module across 44 datasets.

Dataset	Object	All	Train.	Val.	Test.
HaN-Seg-CT		85,030			85,030
	Arytenoid	460			460
	Brain stem	4,935			4,935
	Bone mandible	7,390			7,390
	Buccal mucosa	2,955			2,955
	Cochleal	35			35
	Cricopharyngeus	1,800			1,800
	Cochlear	70			70
	Cavityoral	5,695			5,695
	Eyepl	2,500			2,500
	Eyepr	2,495			2,495
	Esophaguss	1,540			1,540
	Eyear	1,075			1,075
	Eyeal	1,015			1,015
	Glndthyroid	4,485			4,485
	Glottis	1,665			1,665
	Glndlacrimalr	450			450
	Glndsubmandl	3,310			3,310
	Glndlacrimall	345			345
	Glndsubmandr	3,360			3,360
	Larynxsg	3,090			3,090
	Left acarotid	4,255			4,255
	Lips	3,795			3,795
	Opticchiasm	360			360
	Opticnrvr	620			620
	Opticnrvl	595			595
	Parotidl	5,440			5,440
	Pituitary	405			405
	Parotidr	5,550			5,550
	Right acarotid	3,455			3,455
	Spinalcord	11,885			11,885
HCC-TACE-Seg-CT		45,665			45,665
	Liver blood vessel	9,875			9,875
	Liver	20,709			20,709
	Liver tumor	15,081			15,081
HECKTOR-PETCT		68,382	53,844	14,538	
	Head & neck lymph nodes	38,113	30,373	7,740	
	Head & neck primary tumor	30,269	23,471	6,798	
INSTANCE-CT		4,482	3,728	754	
	Hematoma	4,482	3,728	754	

Table S2: (Continued, part 3) Detailed data characteristics of the Box2Mask module across 44 datasets.

Dataset	Object	All	Train.	Val.	Test.
KiPA-CT		101,179	78,245	22,934	
	Kidney tumor	22,990	17,410	5,580	
	Kidney	48,140	37,718	10,422	
	Renal artery	12,387	9,508	2,879	
	Renal vein	17,662	13,609	4,053	
KiTS-CT		308,907	232,037	76,870	
	Kidney cyst	20,567	13,570	6,997	
	Kidney tumor	58,049	44,478	13,571	
	Kidney	230,291	173,989	56,302	
LNQ2023-CT		20,847			20,847
	Mediastinal lymph node	20,847			20,847
Lymph Nodes-CT		68,240	55,745	12,495	
	Mediastinal lymph node	68,240	55,745	12,495	
NSCLC Pleural Effusion-CT		348,716	280,775	67,941	
	Effusions	24,575	17,571	7,004	
	Thoracic cavities	324,141	263,204	60,937	
QUBIQ-CT		11,143			11,143
	Pancreatic lesion	2,635			2,635
	Pancreas	8,508			8,508
MSD-Task03 Liver-CT		128,655	99,430	29,225	
	Liver	76,493	57,176	19,317	
	Liver cancer	52,162	42,254	9,908	
MSD-Task06 Lung-CT		7,033	5,385	1,648	
	Lung cancer	7,033	5,385	1,648	
MSD-Task07 Pancreas-CT		50,293	39,913	10,380	
	Pancreas cancer	12,250	9,445	2,805	
	Pancreas	38,043	30,468	7,575	
MSD-Task08 HepaticVessel-CT		31,743	25,860	5,883	
	Hepatic tumour	18,070	14,650	3,420	
	Hepatic vessel	13,673	11,210	2,463	
MSD-Task09 Spleen-CT		5,211	4,051	1,160	
	Spleen	5,211	4,051	1,160	
MSD-Task10 Colon-CT		6,304	4,494	1,810	
	Colon cancer primaries	6,304	4,494	1,810	

Table S2: (Continued, part 4) Detailed data characteristics of the Box2Mask module across 44 datasets.

Dataset	Object	All	Train.	Val.	Test.
TotalSegmentator-CT		12,992,914	10,373,109	2,619,805	
	Autochthonleft	836,226	667,406	168,820	
	Autochthonright	845,499	675,211	170,288	
	Adrenalglandleft	3,516	2,792	724	
	Adrenalglandright	4,257	3,472	785	
	Aorta	585,577	470,003	115,574	
	Brain	47,028	37,330	9,698	
	Colon	485,194	389,829	95,365	
	Claviculaleft	85,636	68,142	17,494	
	Clavicularight	86,348	68,415	17,933	
	Duodenum	126,706	102,338	24,368	
	Esophagus	72,448	60,449	11,999	
	Face	63,740	50,485	13,255	
	Femurleft	153,571	122,466	31,105	
	Femurright	143,222	114,065	29,157	
	Gluteusminimusleft	108,438	86,306	22,132	
	Gluteusmaximusleft	267,032	213,220	53,812	
	Gluteusminimusright	120,611	96,214	24,397	
	Gallbladder	48,037	37,285	10,752	
	Gluteusmediusright	200,405	160,352	40,053	
	Gluteusmaximusright	266,335	212,728	53,607	
	Gluteusmediusleft	184,963	147,562	37,401	
	Hipright	276,351	220,060	56,291	
	Heartventricleright	176,035	140,458	35,577	
	Humerusright	77,629	59,424	18,205	
	Humerusleft	77,609	61,043	16,566	
	Heartatriumright	132,200	105,320	26,880	
	Heartmyocardium	152,321	121,774	30,547	
	Heartventricleleft	119,593	95,703	23,890	
	Heartatriumleft	106,124	84,678	21,446	
	Hipleft	277,792	221,595	56,197	
	Iliacarteryright	17,501	14,720	2,781	
	Iliopsoasleft	408,412	325,430	82,982	
	Iliopsoasright	407,878	325,782	82,096	
	Iliacarteryleft	16,171	13,601	2,570	
	Iliacvenaleft	71,556	57,500	14,056	
	Iliacvenaright	36,837	29,306	7,531	
	Inferiorvenacava	342,786	274,393	68,393	
	Kidneyleft	163,095	130,877	32,218	
	Kidneyright	158,404	127,252	31,152	
	Lungupperlobeleft	385,661	306,795	78,866	
	Lunglowerloberight	309,506	246,832	62,674	
	Liver	304,774	243,787	60,987	
	Lungmiddleloberight	183,311	147,026	36,285	
	Lunglowerlobeleft	342,773	272,951	69,822	
	Lungupperloberight	238,264	187,998	50,266	
	Pancreas	124,617	100,476	24,141	
	Portalveinandsplenicvein	69,122	55,463	13,659	
	Pulmonaryartery	80,559	64,108	16,451	
	Ribleft	302,820	241,426	61,394	
	Ribright	308,710	247,718	60,992	

Table S2: (Continued, part 5) Detailed data characteristics of the Box2Mask module across 44 datasets.

Dataset	Object	All	Train.	Val.	Test.
TotalSegmentator-CT [Continued]					
	Scapularight	236,479	187,909	48,570	
	Smallbowel	322,701	258,590	64,111	
	Sacrum	136,445	108,720	27,725	
	Spleen	178,344	143,016	35,328	
	Scapulaleft	236,708	188,462	48,246	
	Stomach	200,349	161,355	38,994	
	Trachea	168,728	133,546	35,182	
	Urinarybladder	75,390	59,482	15,908	
	Vertebraet	640,910	509,930	130,980	
	Vertebrael	317,057	253,617	63,440	
	Vertebraec	76,603	58,916	17,687	
WORD-CT		353,643			353,643
	Adrenal	10,139			10,139
	Bladder	12,993			12,993
	Colon	59,379			59,379
	Duodenum	18,258			18,258
	Esophagus	13,902			13,902
	Gallbladder	6,462			6,462
	Head of femur	34,281			34,281
	Intestine	49,960			49,960
	Liver	30,040			30,040
	Left kidney	21,325			21,325
	Pancreas	15,700			15,700
	Right kidney	20,393			20,393
	Rectum	18,520			18,520
	Spleen	18,643			18,643
	Stomach	23,648			23,648
ACDC-MR		37,053			37,053
	Left ventricle	12,625			12,625
	Myocardium	13,642			13,642
	Right ventricle	10,786			10,786
AMOS-MR		198,236	164,956	33,280	
	Arota	40,452	33,442	7,010	
	Bladder	275	275		
	Duodenum	13,254	11,053	2,201	
	Esophagus	3,305	2,880	425	
	Gallbladder	5,786	5,056	730	
	Left kidney	17,263	14,145	3,118	
	Liver	26,027	21,499	4,528	
	Left adrenal gland	1,591	1,441	150	
	Prostate or uterus	185	185		
	Pancreas	13,596	11,388	2,208	
	Postcava	27,092	22,613	4,479	
	Right kidney	17,267	14,107	3,160	
	Right adrenal gland	1,120	975	145	
	Spleen	16,384	13,574	2,810	
	Stomach	14,639	12,323	2,316	

Table S2: (Continued, part 6) Detailed data characteristics of the Box2Mask module across 44 datasets.

Dataset	Object	All	Train.	Val.	Test.
ATLAS-R2.0-MR		71,081	54,559	16,522	
	Brain stroke	71,081	54,559	16,522	
BraTS-MR		1,736,276	1,383,089	353,187	
	Enhancing brain tumor	657,162	522,206	134,956	
	Brain tumor peritumoral edema	727,826	576,256	151,570	
	Nonenhancing brain tumor core	351,288	284,627	66,661	
CHAOS-MR		11,491			11,491
	Left kidney	2,435			2,435
	Liver	4,150			4,150
	Right kidney	2,541			2,541
	Spleen	2,365			2,365
ISLES-MR		10,660	8,226	2,434	
	Ischemic stroke lesion	10,660	8,226	2,434	
MnM2-MR		77,936	62,170	15,766	
	Left ventricle	27,244	21,722	5,522	
	Myocardium	27,206	21,733	5,473	
	Right ventricle	23,486	18,715	4,771	
NCI-ISBI-MR		8,213	6,968	1,245	
	Prostate central gland	5,056	4,297	759	
	Prostate peripheral	3,157	2,671	486	
PI-CAI-MR		17,723	14,085	3,638	
	Prostate cancer	17,723	14,085	3,638	
PROMISE-MR		7,356	5,636	1,720	
	Prostate	7,356	5,636	1,720	
Qin-Prostate-Repeatability-MR		5,113	4,369	744	
	Prostate gland peripheral zone	1,882	1,576	306	
	Prostate suspected tumor	1,149	965	184	
	Prostate gland	393	358	35	
	Prostate	1,689	1,470	219	
Spine-MR		9,265	6,690	2,575	
	Sacral spine	1,250	845	405	
	Lumbar spine	2,685	1,920	765	
	Thoracic spine	5,330	3,925	1,405	
MSD-Task01 BrainTumour-MR		215,049	168,586	46,463	
	Edema	98,320	77,235	21,085	
	Enhancing tumor	83,366	65,587	17,779	
	Nonenhancing tumor	33,363	25,764	7,599	
MSD-Task02 Heart-MR		6,185	4,730	1,455	
	Left atrium	6,185	4,730	1,455	

Table S2: (Continued, part 7) Detailed data characteristics of the Box2Mask module across 44 datasets.

Dataset	Object	All	Train.	Val.	Test.
MSD-Task04 Hippocampus-MR		18,230	14,214	4,016	
	Anterior	10,381	8,051	2,330	
	Posterior	7,849	6,163	1,686	
MSD-Task05 Prostate-MR		6,369	5,122	1,247	
	Peripheral zone	1,979	1,572	407	
	Transitional zone	4,390	3,550	840	
WMH-MR		18,748	15,213	3,535	
	Other pathology	2,348	1,821	527	
	White matter hyperintensities	16,400	13,392	3,008	

Table S3: Detailed data characteristics of the PropMask module across 44 datasets.

The table includes segmentation objects contained in each dataset, the number of tasks used, and the distribution of these tasks across training (Train.), validation (Val.), and testing (Test.) sets.

Dataset	Object	All	Train.	Val.	Test.
AbdomenCT-1K-CT		88,811	69,492	19,319	
	Kidneys	30,840	24,101	6,739	
	Liver	19,871	15,599	4,272	
	Pancreas	19,058	14,896	4,162	
	Spleen	19,042	14,896	4,146	
Adrenal-ACC-Ki67-Seg-CT		903			903
	Adrenocorticalcarcinoma	903			903
AMOS-CT		69,699	55,109	14,590	
	Arota	5,981	4,708	1,273	
	Bladder	3,523	2,818	705	
	Duodenum	5,465	4,301	1,164	
	Esophagus	4,448	3,518	930	
	Gallbladder	3,088	2,433	655	
	Leftkidney	5,818	4,592	1,226	
	Liver	5,947	4,674	1,273	
	Leftadrenalgland	2,557	2,054	503	
	Pancreas	5,104	4,045	1,059	
	Postcava	5,970	4,695	1,275	
	Prostateoruterus	3,191	2,572	619	
	Rightkidney	5,767	4,534	1,233	
	Rightadrenalgland	2,118	1,676	442	
	Stomach	5,411	4,265	1,146	
	Spleen	5,311	4,224	1,087	
AutoPET-PETCT		1,094	834	260	
	Lesion	1,094	834	260	
AutoPET-CT		1,138	853	285	
	Lesion	1,138	853	285	
CHAOS-CT		400			400
	Liver	400			400
COVID-19 Seg. Challenge-CT		4,262	3,410	852	
	Covid19infections	4,262	3,410	852	
COVID-19-CT-Seg-CT		1,979	1,692	287	
	Covid19infections	1,123	956	167	
	Leftlung	442	382	60	
	Rightlung	414	354	60	

Table S3: (Continued, part 2) Detailed data characteristics of the PropMask module across 44 datasets.

Dataset	Object	All	Train.	Val.	Test.
HaN-Seg-CT		12,097			12,097
	Arytenoid	92			92
	Buccalmucosa	570			570
	Bonemandible	808			808
	Brainstem	798			798
	Cricopharyngeus	360			360
	Cavityoral	828			828
	Cochleal	7			7
	Cochlear	14			14
	Eyepr	499			499
	Eyeal	203			203
	Esophaguss	309			309
	Eyepl	501			501
	Eyear	215			215
	Glndsubmandr	657			657
	Glottis	333			333
	Glndsubmandl	645			645
	Glndthyroid	739			739
	Glndlacrimall	69			69
	Glndlacrimalr	90			90
	Larynxsg	595			595
	Left acarotid	173			173
	Lips	736			736
	Opticchiasm	72			72
	Opticnrvr	124			124
	Opticnrvl	119			119
	Parotidr	810			810
	Parotidl	800			800
	Pituitary	81			81
	Right acarotid	170			170
	Spinalcord	680			680
HCC-TACE-Seg-CT		5,591			5,591
	Liver blood vessel	1,921			1,921
	Liver	1,940			1,940
	Liver tumor	1,730			1,730
HECKTOR-PETCT		10,595	8,775	1,820	
	Head & neck lymph nodes	5,470	4,589	881	
	Head & neck primary tumor	5,125	4,186	939	
INSTANCE-CT		852	710	142	
	Hematoma	852	710	142	

Table S3: (Continued, part 3) Detailed data characteristics of the PropMask module across 44 datasets.

Dataset	Object	All	Train.	Val.	Test.
KiPA-CT		4,470	3,494	976	
	Kidney tumor	1,366	1,067	299	
	Kidney	1,371	1,074	297	
	Renal artery	568	423	145	
	Renal vein	1,165	930	235	
KiTS-CT		23,586	19,276	4,310	
	Kidney cyst	2,685	2,238	447	
	Kidney tumor	6,694	5,505	1,189	
	Kidney	14,207	11,533	2,674	
LNQ2023-CT		3,851			3,851
	Mediastinal lymph node	3,851			3,851
Lymph Nodes-CT		13,722	10,842	2,880	
	Mediastinal lymph nodes	13,722	10,842	2,880	
NSCLC Pleural Effusion-CT		16,219	12,798	3,421	
	Effusions	2,167	1,715	452	
	Thoracic cavities	14,052	11,083	2,969	
QUBIQ-CT		1,172			1,172
	Pancreas	879			879
	Pancreatic lesion	293			293
MSD-Task03 Liver-CT		6,652	5,168	1,484	
	Liver	2,599	2,083	516	
	Liver cancer	4,053	3,085	968	
MSD-Task06 Lung-CT		870	634	236	
	Lung cancer	870	634	236	
MSD-Task07 Pancreas-CT		7,657	5,754	1,903	
	Pancreas cancer	2,311	1,693	618	
	Pancreas	5,346	4,061	1,285	
MSD-Task08 HepaticVessel-CT		6,348	4,912	1,436	
	Hepatic tumour	2,792	2,196	596	
	Hepatic vessel	3,556	2,716	840	
MSD-Task09 Spleen-CT		724	544	180	
	Spleen	724	544	180	
MSD-Task10 Colon-CT		1,233	954	279	
	Colon cancer primaries	1,233	954	279	

Table S3: (Continued, part 4) Detailed data characteristics of the PropMask module across 44 datasets.

Dataset	Object	All	Train.	Val.	Test.
TotalSegmentator-CT		1,007,039	800,350	206,689	
	Autochthonleft	16,345	13,038	3,307	
	Aorta	15,236	12,194	3,042	
	Autochthonright	16,511	13,171	3,340	
	Adrenalglandleft	537	426	111	
	Adrenalglandright	721	581	140	
	Brain	3,000	2,397	603	
	Claviculaleft	8,135	6,480	1,655	
	Colon	12,964	10,345	2,619	
	Clavicularight	8,133	6,452	1,681	
	Duodenum	9,406	7,595	1,811	
	Esophagus	3,127	2,529	598	
	Femurleft	8,388	6,634	1,754	
	Face	4,177	3,271	906	
	Femurright	8,250	6,524	1,726	
	Gluteusmediusleft	8,953	7,023	1,930	
	Gallbladder	6,520	5,276	1,244	
	Gluteusminimusleft	8,130	6,371	1,759	
	Gluteusmediusright	8,799	6,911	1,888	
	Gluteusminimusright	8,030	6,320	1,710	
	Gluteusmaximusleft	8,889	6,985	1,904	
	Gluteusmaximusright	8,722	6,878	1,844	
	Hipright	9,316	7,367	1,949	
	Humerusleft	7,072	5,622	1,450	
	Heartmyocardium	13,246	10,552	2,694	
	Heartatriumleft	10,984	8,784	2,200	
	Heartventricleleft	12,470	9,942	2,528	
	Heartventricleright	12,960	10,321	2,639	
	Humerusright	7,261	5,707	1,554	
	Heartatriumright	12,184	9,720	2,464	
	Hipleft	9,440	7,428	2,012	
	Iliopsoasright	9,448	7,513	1,935	
	Iliacarteryright	679	543	136	
	Iliacvenaleft	2,826	2,204	622	
	Iliacvenaright	1,454	1,146	308	
	Inferiorvenacava	12,079	9,670	2,409	
	Iliopsoasleft	9,684	7,715	1,969	
	Iliacarteryleft	628	473	155	
	Kidneyleft	10,839	8,731	2,108	
	Kidneyright	10,662	8,573	2,089	
	Lunglowerlobeleft	14,179	11,288	2,891	
	Lungmiddleloberight	12,616	10,077	2,539	
	Lungupperloberight	11,825	9,298	2,527	
	Liver	14,281	11,458	2,823	
	Lunglowerloberight	13,888	11,109	2,779	
	Lungupperlobeleft	14,526	11,531	2,995	
	Pulmonaryartery	9,338	7,366	1,972	
	Portalveinandsplenicvein	7,781	6,313	1,468	
	Pancreas	10,266	8,260	2,006	
	Ribright	21,780	17,382	4,398	
	Ribleft	21,162	16,788	4,374	

Table S3: (Continued, part 5) Detailed data characteristics of the PropMask module across 44 datasets.

Dataset	Object	All	Train.	Val.	Test.
TotalSegmentator-CT [Continued]					
	Stomach	13,329	10,735	2,594	
	Scapularight	9,692	7,608	2,084	
	Spleen	12,998	10,456	2,542	
	Sacrum	7,103	5,592	1,511	
	Scapulaleft	9,694	7,631	2,063	
	Smallbowel	11,198	8,965	2,233	
	Trachea	9,266	7,174	2,092	
	Urinarybladder	8,279	6,557	1,722	
	Vertebrael	149,211	118,450	30,761	
	Vertebraet	149,211	118,450	30,761	
	Vertebraec	149,211	118,450	30,761	
WORD-CT		36,076			36,076
	Adrenal	1,477			1,477
	Bladder	2,107			2,107
	Colon	2,940			2,940
	Duodenum	2,343			2,343
	Esophagus	1,915			1,915
	Gallbladder	1,282			1,282
	Head of femur	4,783			4,783
	Intestine	2,590			2,590
	Liver	2,388			2,388
	Left kidney	2,392			2,392
	Pancreas	2,330			2,330
	Rectum	2,359			2,359
	Right kidney	2,385			2,385
	Spleen	2,387			2,387
	Stomach	2,398			2,398
ACDC-MR		4,109			4,109
	Left ventricle	1,398			1,398
	Myocardium	1,439			1,439
	Right ventricle	1,272			1,272
AMOS-MR		11,342	9,627	1,715	
	Arota	985	842	143	
	Bladder	40	40		
	Duodenum	1,035	891	144	
	Esophagus	304	250	54	
	Gallbladder	770	635	135	
	Leftkidney	1,161	986	175	
	Liver	1,170	996	174	
	Left adrenal gland	231	192	39	
	Postcava	861	730	131	
	Pancreas	1,124	952	172	
	Prostate or uterus	19	19		
	Right kidney	1,156	986	170	
	Right adrenal gland	192	169	23	
	Spleen	1,132	955	177	
	Stomach	1,162	984	178	

Table S3: (Continued, part 6) Detailed data characteristics of the PropMask module across 44 datasets.

Dataset	Object	All	Train.	Val.	Test.
ATLAS-R2.0-MR		6,151	4,960	1,191	
	Brain stroke	6,151	4,960	1,191	
BraTS-MR		223,716	182,668	41,048	
	Enhancing tumor	78,468	64,091	14,377	
	Brain tumor peritumoral edema	87,546	71,311	16,235	
	Nonenhancing tumor core	57,702	47,266	10,436	
CHAOS-MR		2,207			2,207
	Left kidney	486			486
	Liver	746			746
	Right kidney	508			508
	Spleen	467			467
ISLES		1,415	1,076	339	
	Ischemic stroke lesion	1,415	1,076	339	
MnM2-MR		8,894	6,892	2,002	
	Left ventricle	3,055	2,369	686	
	Myocardium	3,163	2,455	708	
	Right ventricle	2,676	2,068	608	
NCI-ISBI-MR		1,806	1,496	310	
	Prostate central gland	1,031	859	172	
	Prostate peripheral	775	637	138	
PI-CAI-MR		3,575	2,812	763	
	Prostate cancer	3,575	2,812	763	
PROMISE-MR		894	817	77	
	Prostate	894	817	77	
Qin-Prostate-Repeatability-MR		1,063	824	239	
	Prostate gland peripheral zone	404	321	83	
	Prostate suspected tumor	238	148	90	
	Prostate gland	80	62	18	
	Prostate	341	293	48	
Spine-MR		654	507	147	
	Sacral spine	212	163	49	
	Lumbar spine	249	196	53	
	Thoracic spine	193	148	45	
MSD-Task01 BrainTumour-MR		42,817	33,319	9,498	
	Edema	17,038	13,229	3,809	
	Enhancing tumor	13,518	10,520	2,998	
	Nonenhancing tumor	12,261	9,570	2,691	
MSD-Task02 Heart-MR		373	245	128	
	Left atrium	373	245	128	

Table S3: (Continued, part 7) Detailed data characteristics of the PropMask module across 44 datasets.

Dataset	Object	All	Train.	Val.	Test.
MSD-Task04 Hippocampus-MR		3,843	3,000	843	
	Anterior	2,215	1,721	494	
	Posterior	1,628	1,279	349	
MSD-Task05 Prostate-MR		1,482	1,262	220	
	Peripheral zone	602	516	86	
	Transitional zone	880	746	134	
WMH-MR		2,912	2,370	542	
	Other pathology	458	322	136	
	White matter hyperintensities	2,454	2,048	406	
GlomSeg-microCT					4,782
	Glomerulus				4,782

Table S4: Data fingerprints across 44 medical imaging datasets.

Values in parentheses represent the interquartile range (IQR). 'Size anisotropy' is defined as the ratio of the smallest to the largest dimension in the 3D scan data. 'Spacing anisotropy' refers to the ratio of the smallest to the largest spacing in the dataset. 'Object types' indicates the number of distinct annotated segmentation object categories within each dataset.

Dataset	Number of 3D scans	Number of voxels	Size anisotropy	Spacing anisotropy	Object types	Modality
AbdomenCT-1K	1000	2.7001×10^7 (2.3069×10^7, 5.6623×10^7)	0.2012 (0.1719, 0.4204)	0.3221 (0.2295, 0.5845)	4	CT
AMOS-CT	300	4.8955×10^7 (2.9360×10^7, 5.7934×10^7)	0.1885 (0.1328, 0.3389)	0.1564 (0.1239, 0.2830)	15	CT
AutoPET-PETCT	238	5.2160×10^7 (4.5440×10^7, 5.2280×10^7)	0.7477 (0.7100, 0.8150)	0.6788 (0.6788, 0.6788)	1	PET-CT
AutoPET-CT	238	5.2160×10^7 (4.5440×10^7, 5.2280×10^7)	0.7477 (0.7100, 0.8150)	0.6788 (0.6788, 0.6788)	1	CT
COVID-19 Seg. Challenge	199	1.6515×10^7 (1.5204×10^7, 1.8088×10^7)	0.1230 (0.1133, 0.1348)	0.1562 (0.1481, 0.1728)	1	CT
COVID-19-CT-Seg	20	5.2429×10^7 (1.7860×10^7, 7.2090×10^7)	0.3906 (0.0714, 0.5371)	0.4733 (0.1139, 0.6836)	3	CT
HECKTOR	476	3.5127×10^7 (3.3292×10^7, 7.6022×10^7)	0.2617 (0.2480, 0.5664)	0.2986 (0.2986, 0.5859)	2	PET-CT
INSTANCE	100	7.6022×10^6 (7.3400×10^6, 8.1920×10^6)	0.0566 (0.0547, 0.0610)	0.0906 (0.0860, 0.0976)	1	CT
KiPA	70	4.4884×10^6 (3.7638×10^6, 5.5272×10^6)	0.7820 (0.7151, 0.8334)	1.0000 (1.0000, 1.0000)	4	CT
KiTS	489	2.7263×10^7 (2.2020×10^7, 5.5312×10^7)	0.2031 (0.1641, 0.4102)	0.2327 (0.1562, 0.3877)	3	CT
Lymph Nodes	176	1.6987×10^8 (1.5231×10^8, 1.7859×10^8)	0.7877 (0.7488, 0.8663)	0.7812 (0.7031, 0.8984)	1	CT
NSCLC Pleural Effusion	655	3.1982×10^7 (2.7001×10^7, 3.5127×10^7)	0.2383 (0.2012, 0.2617)	0.3255 (0.3255, 0.3255)	2	CT
MSD-Task03 Liver	131	1.1325×10^8 (4.9807×10^7, 1.7931×10^8)	0.6088 (0.3711, 0.7830)	0.8296 (0.4220, 0.9724)	2	CT
MSD-Task06 Lung	63	6.6060×10^7 (6.3046×10^7, 7.8250×10^7)	0.4922 (0.4697, 0.5830)	0.6588 (0.6222, 0.7099)	1	CT

Table S4: (Continued, part 1) Data fingerprints across 44 medical imaging datasets.

Dataset	Number of 3D scans	Number of voxels	Size anisotropy	Spacing anisotropy	Object types	Modality
MSD-Task07 Pancreas	281	2.4379×10^{7} $(2.1758 \times 10^{7}, 2.7001 \times 10^{7})$	0.1816 (0.1621, 0.2012)	0.3125 (0.2773, 0.3516)	2	CT
MSD-Task08 HepaticVessel	303	1.2845×10^{7} $(1.0748 \times 10^{7}, 2.4510 \times 10^{7})$	0.0957 (0.0801, 0.1826)	0.1719 (0.1518, 0.2883)	2	CT
MSD-Task09 Spleen	41	2.3593×10^{7} $(1.5729 \times 10^{7}, 2.7001 \times 10^{7})$	0.1758 (0.1172, 0.2012)	0.1777 (0.1500, 0.2402)	1	CT
MSD-Task10 Colon	126	2.4904×10^{7} $(2.2610 \times 10^{7}, 2.9557 \times 10^{7})$	0.1855 (0.1685, 0.2202)	0.1641 (0.1465, 0.1872)	1	CT
Total Segmentator	111592	1.3268×10^{7} $(6.5279 \times 10^{6}, 2.3925 \times 10^{7})$	0.7104 (0.5720, 0.8562)	1.0000 (1.0000, 1.0000)	61	CT
AMOS-MR	60	6.6688×10^{6} $(5.9904 \times 10^{6}, 1.9409 \times 10^{7})$	0.2250 (0.1875, 0.3258)	0.3958 (0.3958, 0.7457)	15	MR
ATLAS-R2.0	655	8.6753×10^{6} $(8.6753 \times 10^{6}, 8.6753 \times 10^{6})$	0.8112 (0.8112, 0.8112)	1.0000 (1.0000, 1.0000)	1	MR
BraTS	5004	8.9280×10^{6} $(8.9280 \times 10^{6}, 8.9280 \times 10^{6})$	0.6458 (0.6458, 0.6458)	1.0000 (1.0000, 1.0000)	3	MR
ISLES	500	9.1571×10^{5} $(9.0317 \times 10^{5}, 9.1571 \times 10^{5})$	0.6429 (0.6429, 0.6518)	1.0000 (1.0000, 1.0000)	1	MR
MnM2	702	6.5856×10^{5} $(5.6448 \times 10^{5}, 1.0240 \times 10^{6})$	0.0430 (0.0312, 0.0500)	0.1234 (0.1182, 0.1476)	3	MR
NCI-ISBI	79	3.0966×10^{6} $(2.0480 \times 10^{6}, 5.1200 \times 10^{6})$	0.0703 (0.0625, 0.0800)	0.1500 (0.1333, 0.1736)	2	MR
PI-CAI	2912	3.6864×10^{6} $(3.0966 \times 10^{6}, 8.6016 \times 10^{6})$	0.0495 (0.0297, 0.0595)	0.1667 (0.0969, 0.1667)	1	MR
PROMISE	100	2.4576×10^{6} $(2.0480 \times 10^{6}, 7.0124 \times 10^{6})$	0.0711 (0.0625, 0.0750)	0.1736 (0.1184, 0.1736)	1	MR
Qin-Prostate-Repeatability	165	6.8157×10^{6} $(1.3107 \times 10^{6}, 7.8643 \times 10^{6})$	0.0625 (0.0586, 0.0781)	0.1758 (0.1042, 0.2188)	1	MR
Spine	163	4.6162×10^{6} $(3.9322 \times 10^{6}, 6.5536 \times 10^{6})$	0.0293 (0.0250, 0.0488)	0.1758 (0.1299, 0.1758)	1	MR

Table S4: (Continued, part 2) Data fingerprints across 44 medical imaging datasets.

Dataset	Number of 3D scans	Number of voxels	Size anisotropy	Spacing anisotropy	Object types	Modality
MSD-Task01 BrainTumour	1936	8.9280×10^{6} (8.9280×10^{6}, 8.9280×10^{6})	0.6458 (0.6458, 0.6458)	1.0000 (1.0000, 1.0000)	3	MR
MSD-Task02 Heart	20	1.1776×10^{7} (1.1238×10^{7}, 1.2288×10^{7})	0.3594 (0.3430, 0.3750)	0.9124 (0.9124, 0.9124)	1	MR
MSD-Task04 Hippocampus	260	6.2632×10^{4} (5.8252×10^{4}, 6.7353×10^{4})	0.6804 (0.6305, 0.7255)	1.0000 (1.0000, 1.0000)	2	MR
MSD-Task05 Prostate	64	2.0480×10^{6} (1.6005×10^{6}, 2.0480×10^{6})	0.0625 (0.0617, 0.0625)	0.1736 (0.1736, 0.1736)	2	MR
WMH	340	2.8047×10^{6} (2.7648×10^{6}, 2.8508×10^{6})	0.2000 (0.1875, 0.3242)	0.3255 (0.3194, 0.3333)	2	MR
Adrenal-ACC-Ki67-Seg	52	3.0409×10^{7} (2.3790×10^{7}, 4.7055×10^{7})	0.2266 (0.1772, 0.3506)	0.2533 (0.1456, 0.4248)	1	CT
CHAOS-CT	20	2.8967×10^{7} (2.4904×10^{7}, 5.7475×10^{7})	0.2158 (0.1855, 0.4282)	0.4551 (0.4245, 0.6157)	1	CT
HaN-Seg-CT	1259	1.9608×10^{8} (1.3841×10^{8}, 2.1181×10^{8})	0.1953 (0.1777, 0.2090)	0.3110 (0.2612, 0.3389)	30	CT
HCC-TACE-Seg	101	2.2807×10^{7} (2.0185×10^{7}, 2.5952×10^{7})	0.1699 (0.1504, 0.1934)	0.3125 (0.2812, 0.3281)	3	CT
LNQ2023	354	2.9098×10^{7} (2.4642×10^{7}, 3.3751×10^{7})	0.2168 (0.1836, 0.2515)	1.0000 (1.0000, 1.0000)	1	CT
QUBIQ-CT	192	1.3894×10^{7} (1.2255×10^{7}, 2.3593×10^{7})	0.1035 (0.0913, 0.1758)	1.0000 (1.0000, 1.0000)	2	CT
WORD	120	5.2691×10^{7} (4.7120×10^{7}, 5.7934×10^{7})	0.3926 (0.3511, 0.4316)	0.3255 (0.3255, 0.3255)	16	CT
ACDC	300	5.1814×10^{5} (4.4237×10^{5}, 5.6448×10^{5})	0.0391 (0.0312, 0.0391)	0.1562 (0.1367, 0.1680)	3	MR
CHAOS-MR	20	2.2610×10^{6} (2.1089×10^{6}, 2.4883×10^{6})	0.1176 (0.1042, 0.1338)	0.1888 (0.1601, 0.2260)	4	MR
GlomSeg-microCT	15	1.3107×10^{8} (1.3107×10^{8}, 1.3107×10^{8})	0.9766 (0.9766, 0.9766)	1.0000 (1.0000, 1.0000)	1	micro-CT

Table S5: Performance comparison across 44 datasets.

This table presents a detailed comparative analysis of the performance of two popular algorithms, MedSAM and SegVol, alongside our proposed methods, PropSAM-2DBox and PropSAM-2DMask, across 44 datasets. For internally validated datasets, performance metrics are reported based on the validation sets. For externally validated datasets (indicated by), performance is reported based on their respective external datasets.

Dataset & objects	MedSAM [1]	SegVol [2]	PropSAM-2DBox	PropSAM-2DMask
AbdomenCT-1K	0.552	0.813	0.891	0.905
- Kidneys	0.660	0.912	0.945	0.950
- Liver	0.608	0.838	0.962	0.963
- Pancreas	0.737	0.842	0.949	0.950
- Spleen	0.203	0.658	0.709	0.758
*Adrenal-ACC-Ki67-Seg	0.828	0.793	0.874	0.891
- Adrenocortical carcinoma	0.828	0.793	0.874	0.891
AMOS-CT	0.422	0.540	0.716	0.756
- Arota	0.302	0.721	0.618	0.715
- Bladder	0.630	0.583	0.837	0.857
- Duodenum	0.138	0.402	0.556	0.638
- Esophagus	0.258	0.318	0.407	0.457
- Gallbladder	0.643	0.437	0.804	0.795
- Left adrenal gland	0.107	0.150	0.563	0.650
- Left kidney	0.654	0.722	0.934	0.938
- Liver	0.589	0.865	0.945	0.956
- Pancreas	0.095	0.394	0.649	0.706
- Postcava	0.175	0.652	0.387	0.503
- Prostate or uterus	0.717	0.566	0.792	0.771
- Right adrenal gland	0.240	0.256	0.594	0.684
- Right kidney	0.642	0.702	0.937	0.926
- Spleen	0.691	0.719	0.933	0.949
- Stomach	0.452	0.614	0.785	0.794
AutoPET-PETCT	0.493	0.189	0.669	0.755
- Lesion	0.493	0.189	0.669	0.755
AutoPET-CT	0.375	0.150	0.533	0.544
- Lesion	0.375	0.150	0.533	0.544
*CHAOS-CT	0.670	0.941	0.956	0.960
- Liver	0.670	0.941	0.956	0.960
COVID-19 Seg. Challenge	0.515	0.284	0.619	0.670
- COVID-19 infections	0.515	0.284	0.619	0.670
COVID-19-CT-Seg	0.441	0.610	0.868	0.843
- COVID-19 infections	0.194	0.320	0.725	0.657
- Left lung	0.649	0.798	0.942	0.936
- Right lung	0.480	0.710	0.937	0.935

Table S5: (Continued, part 2) Performance comparison across 44 datasets.

Dataset & objects	MedSAM [1]	SegVol [2]	PropSAM-2DBox	PropSAM-2DMask
*HaN-Seg-CT	0.364	0.468	0.537	0.622
- Acarotidl	0.386	0.317	0.237	0.287
- Acarotidr	0.376	0.325	0.259	0.334
- Arytenoid	0.399	0.172	0.565	0.557
- Bonemandible	0.418	0.637	0.760	0.784
- Brainstem	0.365	0.722	0.556	0.570
- Buccalmucosa	0.374	0.482	0.471	0.534
- Cavityoral	0.326	0.778	0.691	0.761
- Cochleal	0.309	0.164	0.778	0.779
- Cochlear	0.360	0.167	0.775	0.772
- Cricopharyngeus	0.408	0.531	0.491	0.526
- Esophaguss	0.380	0.5363	0.494	0.505
- Eyeal	0.417	0.507	0.599	0.653
- Eyear	0.334	0.528	0.612	0.680
- Eyepl	0.350	0.696	0.768	0.882
- Eyepr	0.374	0.706	0.779	0.885
- Glndlacrimall	0.374	0.355	0.478	0.600
- Glndlacrimalr	0.333	0.401	0.533	0.579
- Glndsubmandl	0.396	0.640	0.751	0.787
- Glndsubmandr	0.367	0.636	0.717	0.748
- Glndthyroid	0.362	0.583	0.559	0.847
- Glottis	0.392	0.425	0.286	0.446
- Larynxsg	0.450	0.546	0.273	0.432
- Lips	0.326	0.569	0.425	0.488
- Opticchiasm	0.415	0.125	0.352	0.642
- Opticnrvl	0.383	0.202	0.410	0.642
- Opticnrvr	0.342	0.209	0.490	0.652
- Parotidl	0.353	0.670	0.631	0.750
- Parotidr	0.375	0.666	0.620	0.755
- Pituitary	0.436	0.218	0.552	0.545
- Spinalcord	0.383	0.534	0.183	0.241
*HCC-TACE-Seg	0.498	0.540	0.625	0.703
- Liver	0.771	0.809	0.844	0.861
- Liver blood vessel	0.194	0.185	0.271	0.464
- Liver tumor	0.527	0.626	0.761	0.783
HECKTOR	0.513	0.500	0.586	0.616
- Headnecklymphnodes	0.414	0.566	0.625	0.610
- Headneckprimarytumor	0.613	0.435	0.547	0.621
INSTANCE	0.705	0.459	0.793	0.868
- Hematoma	0.705	0.459	0.793	0.868

Table S5: (Continued, part 3) Performance comparison across 44 datasets.

Dataset & objects	MedSAM [1]	SegVol [2]	PropSAM-2DBox	PropSAM-2DMask
KiPA	0.364	0.510	0.560	0.757
- Kidney	0.633	0.826	0.883	0.912
- Kidney tumor	0.501	0.745	0.831	0.860
- Renal artery	0.062	0.176	0.196	0.633
- Renal vein	0.262	0.292	0.329	0.622
KiTS	0.426	0.526	0.810	0.793
- Kidney	0.705	0.675	0.907	0.918
- Kidney cyst	0.327	0.630	0.748	0.665
- Kidney tumor	0.247	0.274	0.776	0.798
*LNQ2023	0.678	0.378	0.750	0.787
- Mediastinal lymph node	0.678	0.378	0.750	0.787
Lymph Nodes	0.236	0.431	0.630	0.632
- Mediastinal lymph node	0.236	0.431	0.630	0.632
NSCLC Pleural Effusion	0.315	0.292	0.693	0.786
- Effusions	0.069	0.059	0.445	0.630
- Thoracic cavities	0.560	0.526	0.942	0.941
*QUBIQ-CT	0.364	0.519	0.665	0.701
- Pancreas	0.290	0.399	0.654	0.700
- Pancreatic lesion	0.438	0.639	0.676	0.703
MSD-Task03 Liver	0.412	0.718	0.831	0.851
- Liver	0.615	0.872	0.938	0.945
- Liver cancer	0.209	0.565	0.723	0.757
MSD-Task06 Lung	0.609	0.692	0.737	0.770
- Lung cancer	0.609	0.692	0.737	0.770
MSD-Task07 Pancreas	0.360	0.720	0.698	0.718
- Pancreas	0.250	0.707	0.650	0.682
- Pancreas cancer	0.471	0.733	0.745	0.754
MSD-Task08 HepaticVessel	0.422	0.551	0.525	0.677
- Hepatictumour	0.706	0.718	0.750	0.779
- Hepaticvessel	0.138	0.384	0.301	0.575
MSD-Task09 Spleen	0.667	0.874	0.935	0.943
- Spleen	0.667	0.874	0.935	0.943
MSD-Task10 Colon	0.570	0.606	0.690	0.763
- Colon cancer primaries	0.570	0.606	0.690	0.763

Table S5: (Continued, part 4) Performance comparison across 44 datasets.

Dataset & objects	MedSAM [1]	SegVol [2]	PropSAM-2DBox	PropSAM-2DMask
TotalSegmentor	0.353	0.501	0.688	0.728
- Adrenalglandleft	0.284	0.756	0.482	0.551
- Adrenalglandright	0.361	0.398	0.491	0.576
- Aorta	0.414	0.766	0.702	0.756
- Autochthonleft	0.413	0.763	0.667	0.699
- Autochthonright	0.717	0.805	0.658	0.696
- Brain	0.256	0.455	0.874	0.882
- Claviculaleft	0.250	0.422	0.595	0.619
- Clavicularight	0.246	0.429	0.547	0.591
- Colon	0.378	0.490	0.492	0.665
- Duodenum	0.185	0.555	0.571	0.638
- Esophagus	0.658	0.715	0.288	0.316
- Face	0.580	0.734	0.835	0.845
- Femurleft	0.623	0.771	0.851	0.865
- Femurright	0.537	0.544	0.889	0.905
- Gallbladder	0.670	0.796	0.727	0.752
- Gluteusmaximusleft	0.654	0.806	0.751	0.763
- Gluteusmaximusright	0.563	0.728	0.857	0.868
- Gluteusmediusleft	0.537	0.739	0.750	0.803
- Gluteusmediusright	0.381	0.628	0.852	0.865
- Gluteusminimusleft	0.390	0.634	0.712	0.737
- Gluteusminimusright	0.750	0.748	0.793	0.831
- Heartatriumleft	0.686	0.733	0.824	0.881
- Heartatriumright	0.449	0.495	0.735	0.824
- Heartmyocardium	0.693	0.730	0.489	0.735
- Heartventricleleft	0.712	0.709	0.774	0.863
- Heartventricleright	0.345	0.589	0.786	0.833
- Hipleft	0.340	0.601	0.839	0.844
- Hipright	0.635	0.611	0.888	0.894
- Humerusleft	0.575	0.596	0.840	0.862
- Humerusright	0.145	0.248	0.783	0.810
- Iliacarteryleft	0.143	0.246	0.275	0.312
- Iliacarteryright	0.156	0.284	0.265	0.318
- Iliacvenaleft	0.176	0.335	0.314	0.374
- Iliacvenaright	0.290	0.598	0.252	0.280
- Iliopsoasleft	0.251	0.614	0.694	0.743
- Iliopsoasright	0.517	0.726	0.714	0.765
- Inferiorvenacava	0.651	0.732	0.477	0.516
- Kidneyleft	0.641	0.855	0.897	0.912
- Kidneyright	0.682	0.743	0.887	0.894
- Liver	0.573	0.652	0.933	0.940
- Lunglowerlobeleft	0.558	0.632	0.840	0.849
- Lunglowerloberight	0.581	0.576	0.822	0.832
- Lungmiddleloberight	0.550	0.674	0.721	0.732
- Lungupperlobeleft	0.628	0.627	0.879	0.881
- Lungupperloberight	0.276	0.495	0.759	0.766
- Pancreas	0.182	0.196	0.666	0.730
- Portalveinandsplenicvein	0.309	0.526	0.365	0.589
- Pulmonaryartery	0.185	0.346	0.742	0.771
- Ribleft	0.125	0.279	0.422	0.445
- Ribright	0.133	0.297	0.405	0.437

Table S5: (Continued, part 5) Performance comparison across 44 datasets.

Dataset & objects	MedSAM [1]	SegVol [2]	PropSAM-2DBox	PropSAM-2DMask
TotalSegmentator [Continued]				
- Sacrum	0.246	0.465	0.841	0.863
- Scapulaleft	0.247	0.454	0.698	0.724
- Scapularight	0.354	0.498	0.834	0.858
- Smallbowel	0.689	0.783	0.585	0.653
- Spleen	0.507	0.686	0.899	0.906
- Stomach	0.192	0.659	0.823	0.838
- Trachea	0.723	0.734	0.651	0.683
- Urinarybladder	0.336	0.458	0.826	0.853
- Vertebraec	0.422	0.458	0.621	0.640
- Vertebrael	0.395	0.477	0.814	0.827
- Vertebraet	0.368	0.499	0.695	0.714
*WORD	0.405	0.633	0.640	0.700
- Adrenal	0.142	0.637	0.543	0.627
- Bladder	0.647	0.759	0.898	0.909
- Colon	0.163	0.614	0.369	0.523
- Duodenum	0.088	0.460	0.351	0.470
- Esophagus	0.156	0.486	0.440	0.508
- Gallbladder	0.111	0.528	0.616	0.672
- Head of femur	0.721	0.770	0.702	0.712
- Intestine	0.235	0.354	0.548	0.628
- Left kidney	0.676	0.769	0.897	0.898
- Liver	0.635	0.834	0.892	0.936
- Pancreas	0.089	0.472	0.505	0.624
- Rectum	0.440	0.716	0.589	0.629
- Right kidney	0.484	0.780	0.885	0.903
- Spleen	0.713	0.773	0.868	0.891
- Stomach	0.448	0.465	0.496	0.577
*ACDC	0.373	0.653	0.791	0.860
- Left ventricle	0.781	0.853	0.849	0.892
- Myocardium	0.099	0.373	0.740	0.829
- Right ventricle	0.240	0.732	0.786	0.858
AMOS-MR	0.343	0.553	0.607	0.655
- Arota	0.270	0.733	0.589	0.649
- Duodenum	0.115	0.421	0.432	0.546
- Esophagus	0.177	0.382	0.262	0.363
- Gallbladder	0.703	0.569	0.530	0.491
- Left adrenal gland	0.168	0.325	0.496	0.523
- Left kidney	0.613	0.821	0.898	0.912
- Liver	0.621	0.854	0.870	0.896
- Pancreas	0.107	0.555	0.622	0.676
- Postcava	0.356	0.694	0.379	0.486
- Right adrenal gland	0.275	0.180	0.383	0.439
- Right kidney	0.731	0.781	0.883	0.903
- Spleen	0.678	0.832	0.877	0.910
- Stomach	0.332	0.591	0.674	0.715

Table S5: (Continued, part 6) Performance comparison across 44 datasets.

Dataset & objects	MedSAM [1]	SegVol [2]	PropSAM-2DBox	PropSAM-2DMask
ATLAS-R2.0	0.565	0.394	0.665	0.713
- Brain stroke	0.565	0.394	0.665	0.713
BraTS	0.352	0.352	0.509	0.632
- Enhancing tumor	0.440	0.492	0.626	0.697
- Brain tumor peritumoral edema	0.440	0.533	0.632	0.711
- Nonenhancing tumor core	0.175	0.032	0.268	0.489
*CHAOS-MR	0.570	0.818	0.815	0.807
- Left kidney	0.607	0.831	0.829	0.831
- Liver	0.294	0.799	0.781	0.796
- Right kidney	0.692	0.810	0.800	0.787
- Spleen	0.687	0.831	0.849	0.815
ISLES	0.504	0.500	0.589	0.647
- Ischemic stroke lesion	0.504	0.500	0.589	0.647
MnM2	0.669	0.431	0.696	0.839
- Left ventricle	0.520	0.311	0.611	0.826
- Myocardium	0.732	0.445	0.663	0.811
- Right ventricle	0.755	0.535	0.815	0.878
NCI-ISBI	0.448	0.608	0.768	0.782
- Prostate central gland	0.464	0.793	0.816	0.828
- Prostate peripheral	0.312	0.422	0.719	0.735
PI-CAI	0.811	0.750	0.603	0.656
- Prostate cancer	0.811	0.750	0.603	0.656
PROMISE	0.799	0.748	0.839	0.902
- Prostate	0.799	0.748	0.839	0.902
Qin-Prostate-Repeatability	0.519	0.347	0.522	0.713
- Cervical cancer	0.519	0.347	0.522	0.713
Spine	0.748	0.464	0.880	0.907
- Spine	0.748	0.464	0.880	0.907
MSD-Task01 BrainTumour	0.256	0.347	0.493	0.603
- (FLAIR) Edema	0.456	0.468	0.683	0.738
- (FLAIR) Enhancing tumor	0.131	0.141	0.561	0.627
- (FLAIR) Nonenhancing tumor	0.167	0.458	0.228	0.356
- (T1w) Edema	0.450	0.371	0.536	0.647
- (T1w) Enhancing tumor	0.148	0.142	0.530	0.624
- (T1w) Nonenhancing tumor	0.151	0.432	0.229	0.420
- (T2w) Edema	0.458	0.461	0.620	0.714
- (T2w) Enhancing tumor	0.184	0.453	0.544	0.639
- (T2w) Nonenhancing tumor	0.155	0.115	0.244	0.418
- (T1gd) Edema	0.454	0.304	0.585	0.673
- (T1gd) Enhancing tumor	0.147	0.310	0.733	0.785
- (T1gd) Nonenhancing tumor	0.166	0.506	0.429	0.594
MSD-Task02 Heart	0.649	0.525	0.854	0.863
- Left atrium	0.649	0.525	0.854	0.863

Table S5: (Continued, part 7) Performance comparison across 44 datasets.

Dataset & objects	MedSAM [1]	SegVol [2]	PropSAM-2DBox	PropSAM-2DMask
MSD-Task04 Hippocampus	0.482	0.465	0.597	0.658
- Anterior	0.603	0.613	0.643	0.726
- Posterior	0.361	0.317	0.552	0.591
MSD-Task05 Prostate	0.393	0.528	0.719	0.713
- Peripheral zone	0.351	0.378	0.702	0.639
- Transitional zone	0.434	0.679	0.735	0.788
WMH	0.387	0.013	0.554	0.613
- Other pathology	0.519	0.024	0.666	0.657
- White matter hyperintensities	0.255	0.001	0.443	0.569
*GlomSeg-microCT	0.712	0.080	0.769	0.874
- Glomerulus	0.712	0.080	0.769	0.874

Table S6: Comparison of volumetric segmentation inference times across four models on 44 datasets.

Inference time comparison was conducted on all available (i.e. training, validation, and test) samples. All inference computations were performed using a single NVIDIA A800-SXM4-80GB GPU and eight Intel(R) Xeon(R) Platinum 8358P CPUs at 2.60GHz. Each value represents the average time (in seconds) required to infer a single volumetric segmentation for each dataset. **Bolded values** indicate the fastest inference time, while underlined values denote the second fastest.

ID	Dataset	MedSAM	SegVol	PropSAM-2dbox	PropSAM-2dmask
D01	AbdomenCT-1K	41.8780	12.6614	9.6679	**8.8770**
D02	AMOS-CT	70.2122	27.3987	16.9490	**14.9562**
D03	AutoPET-PETCT	8.3469	4.4771	3.2827	**3.0941**
D04	AutoPET-CT	8.3920	4.4264	3.6519	**3.1835**
D05	COVID-19 Seg. Challenge	6.2545	2.0538	1.5501	**0.9748**
D06	COVID-19-CT-Seg	90.3073	**9.2239**	10.4204	9.7650
D07	HECKTOR	9.7150	9.0745	3.3034	**2.8407**
D08	INSTANCE	1.7371	0.7783	0.4538	**0.2112**
D09	KiPA	43.6810	1.0576	**2.5978**	1.8944
D10	KiTS	63.6858	9.2998	6.7493	**6.2454**
D11	Lymph nodes	25.0051	20.4979	**18.2933**	18.3020
D12	NSCLC Pleural Effusion	14.4148	3.5249	**0.4389**	0.5926
D13	MSD-Task03 Liver	23.5253	19.8104	11.4904	**11.0415**
D14	MSD-Task06 Lung	8.4085	9.3646	1.5383	**1.4343**
D15	MSD-Task07 Pancreas	4.8517	3.6360	1.3514	**1.0263**
D16	MSD-Task08 HepaticVessel	26.4077	2.6219	2.1740	**1.6751**
D17	MSD-Task09 Spleen	4.1238	2.6742	0.7226	**0.5172**
D18	MSD-Task10 Colon	3.2462	3.4183	0.7952	**0.5952**
D19	Total Segmentator	478.6908	29.8210	0.8126	**0.6183**
D20	AMOS-MR	93.5970	8.0589	9.6674	**7.5762**
D21	ATLAS-R2.0	6.6652	0.8115	0.7517	**0.4789**
D22	BraTS	25.7899	**1.1505**	1.8441	1.1781
D23	ISLES	3.6531	**0.1481**	0.6201	0.1482
D24	MnM2	2.3340	0.2789	0.8083	**0.2225**
D25	NCI-ISBI	2.8152	0.5094	0.6535	**0.2490**
D26	PI-CAI	2.0195	0.9505	0.4276	**0.1946**
D27	PROMISE	1.8767	0.6131	0.3881	**0.1723**
D28	Qin-Prostate-Repeatability	4.4264	0.9251	0.3543	**0.1404**
D29	Spine	11.2137	0.6043	0.4192	**0.1789**
D30	MSD-Task01 BrainTumour	43.0633	1.4494	1.8921	**1.2088**
D31	MSD-Task02 Heart	7.6856	1.1063	0.8131	**0.6526**
D32	MSD-Task04 Hippocampus	2.5236	0.1483	0.5340	**0.1308**
D33	MSD-Task05 Prostate	3.7699	0.3993	0.5725	**0.1883**
D34	WMH	13.7550	21.4490	1.8318	**0.5356**
D35	Adrenal-ACC-Ki67-Seg	6.5073	5.0176	1.2562	**1.1041**
D36	CHAOS-CT	15.5163	4.5757	1.9935	**1.8434**
D37	HaN-Seg	229.2443	212.1940	3.2218	**2.8638**
D38	HCC-TACE-Seg	100.6243	4.0337	4.0556	**3.3113**
D39	LNQ2023	3.5902	3.3163	0.9673	**0.6930**
D40	QUBIQ	11.7360	2.5504	0.5800	**0.4209**
D41	WORD	88.1514	33.3395	26.0206	**22.9538**
D42	ACDC	2.6765	0.2596	0.7627	**0.1768**
D43	CHAOS-MR	6.4656	0.7418	1.4681	**0.6870**
D44	GlomSeg-microCT	276.6460	249.8852	**150.009**	161.623

Table S7: Wilcoxon rank sum test for inference time comparison between models

	MedSAM	SegVol	PropSAM-2dbox	PropSAM-2dmask
		P-values		
MedSAM	-	1.3228×10^{-4}	2.3195×10^{-9}	2.9781×10^{-10}
SegVol	-	-	0.0961	0.0060
PropSAM-2dbox	-	-	-	0.1141
PropSAM-2dmask	-	-	-	-

Table S8: Ablation study on the impact of initialization slice deviation on performance in PropSAM.

In this experiment, due to the high computational costs and time constraints, we sampled approximately 60% of the dataset (totaling 26 datasets) and aimed to cover a wide variety of segmented objects for our ablation study. We hypothesized that the optimal slicing plane is the largest cross-section of the segmented object. To simulate initial slice deviation, we randomly offset from this optimal plane in both directions by up to 20%, investigating the impact of starting slice deviation on the performance of PropSAM.

Model	Slice deviation ratio				
	0%	±5%	±10%	±15%	±20%
*Adrenal-ACC-Ki67-Seg					
PropSAM-2DBox	0.874	0.867 (-0.006)	0.852 (-0.022)	0.879 (0.006)	0.859 (-0.014)
PropSAM-2DMask	0.891	0.887 (-0.004)	0.876 (-0.016)	0.875 (-0.016)	0.871 (-0.020)
AutoPET-PETCT					
PropSAM-2DBox	0.669	0.638 (-0.031)	0.653 (-0.017)	0.658 (-0.011)	0.574 (-0.095)
PropSAM-2DMask	0.755	0.740 (-0.015)	0.742 (-0.013)	0.754 (-0.001)	0.725 (-0.029)
*CHAOS-CT					
PropSAM-2DBox	0.956	0.950 (-0.006)	0.938 (-0.018)	0.960 (0.004)	0.821 (-0.135)
PropSAM-2DMask	0.960	0.960 (0.000)	0.947 (-0.013)	0.959 (-0.001)	0.865 (-0.095)
COVID-19 Seg. Challenge					
PropSAM-2DBox	0.619	0.618 (-0.001)	0.636 (0.017)	0.580 (-0.039)	0.584 (-0.035)
PropSAM-2DMask	0.670	0.660 (-0.010)	0.672 (0.002)	0.654 (-0.016)	0.631 (-0.038)
HECKTOR					
PropSAM-2DBox	0.649	0.643 (-0.006)	0.558 (-0.091)	0.604 (-0.046)	0.628 (-0.022)
PropSAM-2DMask	0.677	0.708 (0.030)	0.586 (-0.091)	0.660 (-0.017)	0.670 (-0.007)
INSTANCE					
PropSAM-2DBox	0.793	0.793 (-0.000)	0.793 (0.001)	0.794 (0.001)	0.796 (0.003)
PropSAM-2DMask	0.868	0.868 (-0.000)	0.867 (-0.001)	0.865 (-0.002)	0.851 (-0.017)
*LNQ2023					
PropSAM-2DBox	0.750	0.750 (-0.001)	0.743 (-0.007)	0.736 (-0.015)	0.710 (-0.040)
PropSAM-2DMask	0.787	0.787 (-0.000)	0.782 (-0.005)	0.778 (-0.009)	0.752 (-0.035)
MSD-Task03 Liver					
PropSAM-2DBox	0.831	0.827 (-0.004)	0.843 (0.012)	0.821 (-0.010)	0.737 (-0.093)
PropSAM-2DMask	0.851	0.856 (0.005)	0.875 (0.024)	0.848 (-0.003)	0.793 (-0.058)
MSD-Task06 Lung					
PropSAM-2DBox	0.737	0.718 (-0.018)	0.710 (-0.026)	0.727 (-0.010)	0.634 (-0.103)
PropSAM-2DMask	0.770	0.793 (0.023)	0.770 (0.000)	0.758 (-0.012)	0.705 (-0.065)
MSD-Task07 Pancreas					
PropSAM-2DBox	0.698	0.695 (-0.002)	0.661 (-0.037)	0.659 (-0.039)	0.556 (-0.142)
PropSAM-2DMask	0.718	0.724 (0.006)	0.697 (-0.021)	0.698 (-0.020)	0.599 (-0.119)
MSD-Task08 HepaticVessel					
PropSAM-2DBox	0.525	0.485 (-0.040)	0.461 (-0.064)	0.424 (-0.101)	0.406 (-0.119)
PropSAM-2DMask	0.677	0.673 (-0.004)	0.644 (-0.033)	0.618 (-0.059)	0.573 (-0.104)
MSD-Task09 Spleen					
PropSAM-2DBox	0.935	0.935 (0.000)	0.934 (-0.001)	0.937 (0.002)	0.936 (0.001)
PropSAM-2DMask	0.943	0.943 (-0.000)	0.942 (-0.001)	0.942 (-0.002)	0.941 (-0.002)

Table S8: (Continued, part 1) Ablation study on the impact of initialization slice deviation on performance in PropSAM.

Model	Slice deviation ratio				
	0%	±5%	±10%	±15%	±20%
			MSD-Task10 Colon		
PropSAM-2DBox	0.690	0.692 (0.001)	0.636 (-0.055)	0.678 (-0.013)	0.621 (-0.069)
PropSAM-2DMask	0.763	0.762 (-0.000)	0.741 (-0.021)	0.742 (-0.020)	0.703 (-0.060)
			*WORD		
PropSAM-2DBox	0.640	0.628 (-0.012)	0.616 (-0.024)	0.615 (-0.025)	0.579 (-0.061)
PropSAM-2DMask	0.700	0.706 (0.005)	0.703 (0.003)	0.679 (-0.021)	0.648 (-0.052)
			*ACDC		
PropSAM-2DBox	0.791	0.791 (0.000)	0.788 (-0.004)	0.762 (-0.030)	0.792 (0.001)
PropSAM-2DMask	0.860	0.860 (0.000)	0.860 (0.000)	0.857 (-0.003)	0.864 (0.004)
			AMOS-MR		
PropSAM-2DBox	0.607	0.600 (-0.008)	0.550 (-0.058)	0.551 (-0.057)	0.461 (-0.146)
PropSAM-2DMask	0.655	0.637 (-0.017)	0.600 (-0.054)	0.629 (-0.026)	0.522 (-0.133)
			ATLAS-R2.0		
PropSAM-2DBox	0.665	0.667 (0.002)	0.644 (-0.021)	0.636 (-0.029)	0.604 (-0.061)
PropSAM-2DMask	0.713	0.731 (0.017)	0.726 (0.012)	0.704 (-0.009)	0.670 (-0.043)
			*CHAOS-MR		
PropSAM-2DBox	0.801	0.799 (-0.001)	0.801 (0.000)	0.788 (-0.012)	0.797 (-0.004)
PropSAM-2DMask	0.763	0.763 (0.000)	0.771 (0.008)	0.779 (0.016)	0.770 (0.007)
			ISLES		
PropSAM-2DBox	0.589	0.608 (0.019)	0.612 (0.023)	0.581 (-0.008)	0.559 (-0.030)
PropSAM-2DMask	0.647	0.661 (0.014)	0.668 (0.021)	0.592 (-0.055)	0.599 (-0.048)
			MnM2		
PropSAM-2DBox	0.696	0.696 (0.000)	0.675 (-0.021)	0.612 (-0.084)	0.676 (-0.020)
PropSAM-2DMask	0.839	0.839 (0.000)	0.838 (-0.000)	0.836 (-0.003)	0.835 (-0.003)
			NCI-ISBI		
PropSAM-2DBox	0.768	0.768 (0.000)	0.806 (0.038)	0.795 (0.027)	0.746 (-0.021)
PropSAM-2DMask	0.782	0.782 (0.000)	0.825 (0.044)	0.809 (0.028)	0.813 (0.031)
			PROMISE		
PropSAM-2DBox	0.839	0.839 (0.000)	0.802 (-0.037)	0.864 (0.025)	0.812 (-0.027)
PropSAM-2DMask	0.902	0.902 (-0.000)	0.846 (-0.056)	0.910 (0.008)	0.841 (-0.062)
			Qin-Prostate-Repeatability		
PropSAM-2DBox	0.522	0.522 (0.000)	0.603 (0.081)	0.518 (-0.004)	0.623 (0.101)
PropSAM-2DMask	0.713	0.713 (0.000)	0.719 (0.006)	0.732 (0.019)	0.670 (-0.042)
			Spine		
PropSAM-2DBox	0.880	0.880 (-0.000)	0.864 (-0.016)	0.879 (-0.002)	0.881 (0.001)
PropSAM-2DMask	0.907	0.907 (0.000)	0.907 (-0.000)	0.904 (-0.003)	0.917 (0.010)
			MSD-Task02 Heart		
PropSAM-2DBox	0.854	0.831 (-0.023)	0.749 (-0.106)	0.707 (-0.147)	0.790 (-0.064)
PropSAM-2DMask	0.863	0.815 (-0.048)	0.775 (-0.088)	0.846 (-0.017)	0.833 (-0.029)
			WMH		
PropSAM-2DBox	0.554	0.551 (-0.003)	0.526 (-0.029)	0.460 (-0.095)	0.451 (-0.104)
PropSAM-2DMask	0.613	0.611 (-0.002)	0.591 (-0.022)	0.536 (-0.077)	0.502 (-0.111)

Table S9: Ablation study on the impact of propagation slice thickness on performance in PropSAM.

The dataset used in this experiment is the same as that of Supplementary Table S8. We used four inference thicknesses of 10 mm, 20 mm, 30 mm, and 40 mm to study their impact on the propagation of PropSAM. In this study, we empirically selected 20 mm as our basic setting, which, for most 3D medical scans with a slice thickness of 5 mm, allows the propagation of information across about 4 slices in one direction, thus balancing inference speed and accuracy.

Model	Propagation thickness			
	10 mm	20 mm	30 mm	40 mm
*Adrenal-ACC-Ki67-Seg				
PropSAM-2DBox	0.870 (-0.004)	0.874	0.868 (-0.006)	0.871 (-0.002)
PropSAM-2DMask	0.873 (-0.018)	0.891	0.891 (-0.000)	0.893 (0.002)
AutoPET-PETCT				
PropSAM-2DBox	0.674 (0.005)	0.669	0.658 (-0.011)	0.658 (-0.011)
PropSAM-2DMask	0.739 (-0.016)	0.755	0.733 (-0.022)	0.733 (-0.022)
*CHAOS-CT				
PropSAM-2DBox	0.948 (-0.008)	0.956	0.956 (0.000)	0.956 (0.000)
PropSAM-2DMask	0.954 (-0.006)	0.960	0.960 (0.000)	0.960 (0.000)
COVID-19 Seg. Challenge				
PropSAM-2DBox	0.603 (-0.016)	0.619	0.623 (0.004)	0.629 (0.011)
PropSAM-2DMask	0.626 (-0.044)	0.670	0.672 (0.003)	0.699 (0.030)
HECKTOR				
PropSAM-2DBox	0.698 (0.049)	0.649	0.635 (-0.014)	0.633 (-0.016)
PropSAM-2DMask	0.707 (0.030)	0.677	0.676 (-0.001)	0.668 (-0.009)
INSTANCE				
PropSAM-2DBox	0.787 (-0.006)	0.793	0.792 (-0.001)	0.780 (-0.013)
PropSAM-2DMask	0.859 (-0.009)	0.868	0.851 (-0.017)	0.848 (-0.020)
*LNQ2023				
PropSAM-2DBox	0.750 (0.000)	0.750	0.750 (0.000)	0.750 (0.000)
PropSAM-2DMask	0.787 (0.000)	0.787	0.787 (0.000)	0.787 (0.000)
MSD-Task03 Liver				
PropSAM-2DBox	0.830 (-0.001)	0.831	0.830 (-0.000)	0.830 (-0.001)
PropSAM-2DMask	0.830 (-0.020)	0.851	0.852 (0.001)	0.852 (0.001)
MSD-Task06 Lung				
PropSAM-2DBox	0.736 (-0.001)	0.737	0.737 (0.000)	0.737 (0.000)
PropSAM-2DMask	0.792 (0.023)	0.770	0.770 (0.000)	0.770 (0.000)
MSD-Task07 Pancreas				
PropSAM-2DBox	0.714 (0.016)	0.698	0.703 (0.005)	0.704 (0.006)
PropSAM-2DMask	0.731 (0.013)	0.718	0.729 (0.011)	0.731 (0.013)
MSD-Task08 HepaticVessel				
PropSAM-2DBox	0.525 (-0.001)	0.525	0.515 (-0.011)	0.520 (-0.006)
PropSAM-2DMask	0.662 (-0.015)	0.677	0.677 (0.000)	0.675 (-0.001)
MSD-Task09 Spleen				
PropSAM-2DBox	0.933 (-0.003)	0.935	0.933 (-0.002)	0.932 (-0.003)
PropSAM-2DMask	0.938 (-0.005)	0.943	0.944 (0.001)	0.941 (-0.002)

Table S9: (Continued, part 1) Ablation study on the impact of propagation slice thickness on performance in PropSAM.

Model	Propagation thickness			
	10 mm	20 mm	30 mm	40 mm
MSD-Task10 Colon				
PropSAM-2DBox	0.623 (-0.067)	0.690	0.652 (-0.038)	0.688 (-0.003)
PropSAM-2DMask	0.712 (-0.050)	0.763	0.740 (-0.023)	0.769 (0.006)
*WORD				
PropSAM-2DBox	0.642 (0.003)	0.640	0.639 (-0.001)	0.639 (-0.001)
PropSAM-2DMask	0.683 (-0.017)	0.700	0.698 (-0.002)	0.698 (-0.002)
*ACDC				
PropSAM-2DBox	0.801 (0.010)	0.791	0.788 (-0.004)	0.784 (-0.008)
PropSAM-2DMask	0.852 (-0.008)	0.860	0.864 (0.004)	0.864 (0.004)
AMOS-MR				
PropSAM-2DBox	0.615 (0.008)	0.607	0.600 (-0.008)	0.600 (-0.008)
PropSAM-2DMask	0.648 (-0.006)	0.655	0.647 (-0.007)	0.647 (-0.007)
ATLAS-R2.0				
PropSAM-2DBox	0.665 (0.000)	0.665	0.665 (0.000)	0.665 (0.000)
PropSAM-2DMask	0.713 (0.000)	0.713	0.713 (0.000)	0.713 (0.000)
*CHAOS-MR				
PropSAM-2DBox	0.774 (-0.027)	0.801	0.808 (0.007)	0.831 (0.030)
PropSAM-2DMask	0.735 (-0.028)	0.763	0.789 (0.026)	0.821 (0.058)
ISLES				
PropSAM-2DBox	0.628 (0.039)	0.589	0.592 (0.003)	0.586 (-0.002)
PropSAM-2DMask	0.688 (0.041)	0.647	0.650 (0.003)	0.655 (0.008)
MnM2				
PropSAM-2DBox	0.698 (0.002)	0.696	0.684 (-0.012)	0.681 (-0.015)
PropSAM-2DMask	0.826 (-0.012)	0.839	0.843 (0.005)	0.848 (0.009)
NCI-ISBI				
PropSAM-2DBox	0.801 (0.034)	0.768	0.779 (0.012)	0.799 (0.031)
PropSAM-2DMask	0.780 (-0.001)	0.782	0.809 (0.027)	0.817 (0.036)
PROMISE				
PropSAM-2DBox	0.882 (0.043)	0.839	0.849 (0.010)	0.846 (0.007)
PropSAM-2DMask	0.910 (0.008)	0.902	0.897 (-0.005)	0.905 (0.003)
Qin-Prostate-Repeatability				
PropSAM-2DBox	0.514 (-0.008)	0.522	0.522 (-0.001)	0.522 (-0.001)
PropSAM-2DMask	0.705 (-0.007)	0.713	0.738 (0.025)	0.738 (0.025)
Spine				
PropSAM-2DBox	0.885 (0.004)	0.880	0.869 (-0.012)	0.870 (-0.010)
PropSAM-2DMask	0.903 (-0.003)	0.907	0.897 (-0.010)	0.907 (-0.000)
MSD-Task02 Heart				
PropSAM-2DBox	0.841 (-0.013)	0.854	0.854 (0.000)	0.854 (0.000)
PropSAM-2DMask	0.823 (-0.039)	0.863	0.863 (0.000)	0.863 (0.000)
WMH				
PropSAM-2DBox	0.538 (-0.016)	0.554	0.578 (0.024)	0.545 (-0.009)
PropSAM-2DMask	0.612 (-0.000)	0.613	0.583 (-0.030)	0.594 (-0.019)

Table S10: Generalization analysis of different models on validation sets derived from 10 external datasets.

The table represents each external dataset as a column, where a gray background indicates datasets not encountered during the training phase of the tested models, such as MedSAM, PropSAM-2DBox, and PropSAM-2DMask. The PropSAM-2DMask△ denotes fine-tuning based on the weights of PropSAM-2DMask on datasets highlighted in blue. Performance improvements over the generic model PropSAM-2DMask are marked in red, while declines are shown in blue. PropSAM-2DMask◯ indicates training from scratch using the architecture of PropSAM-2DMask on the training set of datasets marked in blue. Please note that the performance reported here pertains to the validation sets derived from external datasets, as the experiment requires training sets for PropSAM-2DMask△ and PropSAM-2DMask◯ fine-tuning. This is different from Supplementary Table S5, where all data from the external datasets are used for validation purposes.

Dataset IDs									
D35	D36	D37	D38	D39	D40	D41	D42	D43	D44
MedSAM									
0.803	0.679	0.017	0.301	0.675	0.183	0.401	0.432	0.580	0.719
PropSAM-2DBox									
0.821	0.958	0.538	0.633	0.745	0.607	0.654	0.795	0.788	0.754
PropSAM-2DMask									
0.837	0.961	0.619	0.701	0.785	0.635	0.689	0.854	0.751	0.860
PropSAM-2DMask△									
0.879	0.957	0.521	0.576	0.736	0.564	0.652	0.814	0.758	0.654
0.721	0.970	0.564	0.626	0.741	0.589	0.674	0.848	0.774	0.731
0.713	0.926	0.681	0.656	0.711	0.676	0.555	0.840	0.806	0.803
0.832	0.965	0.555	0.729	0.792	0.613	0.635	0.842	0.715	0.460
0.855	0.960	0.601	0.652	0.814	0.621	0.656	0.779	0.850	0.894
0.851	0.949	0.589	0.663	0.775	0.719	0.698	0.818	0.827	0.815
0.741	0.952	0.589	0.665	0.774	0.603	0.679	0.845	0.763	0.865
0.736	0.957	0.572	0.647	0.696	0.593	0.662	0.895	0.821	0.895
0.799	0.962	0.603	0.665	0.776	0.606	0.681	0.839	0.922	0.905
0.876	0.947	0.584	0.701	0.742	0.609	0.668	0.812	0.886	0.937
PropSAM-2DMask◯									
0.801	0.851	0.437	0.401	0.749	0.335	0.347	0.630	0.577	0.105
0.211	0.966	0.129	0.322	0.131	0.109	0.048	0.513	0.142	0.052
0.596	0.901	0.564	0.490	0.648	0.350	0.320	0.784	0.733	0.570
0.718	0.958	0.507	0.680	0.662	0.340	0.478	0.783	0.734	0.264
0.733	0.848	0.453	0.388	0.792	0.371	0.312	0.480	0.515	0.093
0.720	0.033	0.450	0.373	0.724	0.575	0.364	0.534	0.685	0.426
0.459	0.888	0.453	0.518	0.554	0.281	0.519	0.804	0.651	0.473
0.598	0.904	0.352	0.495	0.342	0.360	0.419	0.883	0.749	0.588
0.378	0.957	0.348	0.437	0.236	0.168	0.294	0.804	0.799	0.679
0.576	0.318	0.411	0.320	0.585	0.281	0.263	0.573	0.573	0.886

References

[1] Ma, J. *et al.* Segment anything in medical images. *Nature Communications* **15**, 654 (2024).

[2] Du, Y., Bai, F., Huang, T. & Zhao, B. Segvol: Universal and interactive volumetric medical image segmentation. *arXiv preprint arXiv:2311.13385* (2023).

[3] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241 (Springer, 2015).

[4] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**, 203–211 (2021).

[5] Ulyanov, D., Vedaldi, A. & Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016).

[6] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).

[7] Ma, J. *et al.* Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 6695–6714 (2021).

[8] Ma, J. *et al.* Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis* **82**, 102616 (2022).

[9] Ahmed, A. *et al.* Radiomic mapping model for prediction of ki-67 expression in adrenocortical carcinoma. *Clinical Radiology* **75**, 479–e17 (2020).

[10] Ji, Y. *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).

[11] Gatidis, S. *et al.* A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data* **9**, 601 (2022).

[12] Kavur, A. E. *et al.* Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis* **69**, 101950 (2021).

[13] Roth, H. R. *et al.* Rapid artificial intelligence solutions in a pandemic—the covid-19-20 lung ct lesion segmentation challenge. *Medical image analysis* **82**, 102605 (2022).

[14] Ma, J. *et al.* Toward data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. *Medical physics* **48**, 1197–1210 (2021).

[15] 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**, 1917–1927 (2023).

[16] Clark, K. *et al.* The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging* **26**, 1045–1057 (2013).

[17] Morshid, A. *et al.* A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization. *Radiology: Artificial Intelligence* **1**, e180021 (2019).

[18] Oreiller, V. *et al.* Head and neck tumor segmentation in pet/ct: the hecktor challenge. *Medical image analysis* **77**, 102336 (2022).

[19] Li, X. *et al.* Hematoma expansion context guided intracranial hemorrhage segmentation and uncertainty estimation. *IEEE Journal of Biomedical and Health Informatics* **26**, 1140–1151 (2021).

[20] He, Y. *et al.* Meta grayscale adaptive network for 3d integrated renal structures segmentation. *Medical image analysis* **71**, 102055 (2021).

[21] He, Y. *et al.* Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation. *Medical image analysis* **63**, 101722 (2020).

[22] Heller, N. *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* **67**, 101821 (2021).

[23] Khajavibajestani, R. *et al.* Mediastinal lymph node quantification (lnq): Segmentation of heterogeneous CT data (2023). URL `https://doi.org/10.5281/zenodo.7844666`. [Online].

[24] Roth, H. R. *et al.* A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part I 17*, 520–527 (Springer, 2014).

[25] Kiser, K. J. *et al.* Plethora: Pleural effusion and thoracic cavity segmentations in diseased lungs for benchmarking chest ct processing pipelines. *Medical physics* **47**, 5941–5952 (2020).

[26] Kiser, K. *et al.* Data from the thoracic volume and pleural effusion segmentations in diseased lungs for benchmarking chest ct processing pipelines. the cancer imaging archive (2020).

[27] Becker, A. S. *et al.* Variability of manual segmentation of the prostate in axial t2-weighted mri: a multi-reader study. *European journal of radiology* **121**, 108716 (2019).

[28] Antonelli, M. *et al.* The medical segmentation decathlon. *Nature communications* **13**, 4128 (2022).

[29] Wasserthal, J. *et al.* Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence* **5** (2023).

[30] Luo, X. *et al.* Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *arXiv preprint arXiv:2111.02403* (2021).

[31] Bernard, O. *et al.* Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging* **37**, 2514–2525 (2018).

[32] Liew, S.-L. *et al.* A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific data* **9**, 320 (2022).

[33] Menze, B. H. *et al.* The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging* **34**, 1993–2024 (2014).

[34] Hernandez Petzsche, M. R. *et al.* Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data* **9**, 762 (2022).

[35] Campello, V. M. *et al.* Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging* **40**, 3543–3554 (2021).

[36] Bloch, N. *et al.* NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures. The Cancer Imaging Archive (2015). Available online: `http://doi.org/10.7937/K9/TCIA.2015.zF0vlOPv`.

[37] Saha, A., Hosseinzadeh, M. & Huisman, H. End-to-end prostate cancer detection in bpmri via 3d cnns: effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Medical image analysis* **73**, 102155 (2021).

[38] Litjens, G. *et al.* Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis* **18**, 359–373 (2014).

[39] Fedorov, A. *et al.* Data from qin-prostate-repeatability. The Cancer Imaging Archive (2018). Available online: `https://doi.org/10.7937/K9/TCIA.2018.MR1CKGND`.

[40] Fedorov, A. *et al.* An annotated test-retest collection of prostate multiparametric mri. *Scientific data* **5**, 1–13 (2018).

[41] Zukić, D. *et al.* Robust detection and segmentation for diagnosis of vertebral diseases using routine mr images. In *Computer Graphics Forum*, vol. 33, 190–204 (Wiley Online Library, 2014).

[42] Kuijf, H. J. *et al.* Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging* **38**, 2556–2568 (2019).