MedicoSAM: Robust Improvement of SAM for Medical Imaging

Anwai Archit, Luca Freckmann, Constantin Pape

***Abstract*— Medical image segmentation is an important analysis task in clinical practice and research. Deep learning has massively advanced the field, but current approaches are mostly based on models trained for a specific task. Training such models or adapting them to a new condition is costly due to the need for labeled data. The emergence of vision foundation models, especially Segment Anything Model (SAM), offers a path to universal segmentation for medical images, overcoming these issues. Here, we study how to improve SAM for medical images by comparing different finetuning strategies on a large and diverse dataset. We evaluate the finetuned models on a wide range of interactive and automatic semantic segmentation tasks. We find that performance clearly improves given the correct choice of finetuning strategies. This improvement is especially pronounced for interactive segmentation. Semantic segmentation also benefits, but the advantage over traditional segmentation approaches is inconsistent. Our best model, MedicoSAM, is publicly available. We show that it is compatible with existing tools for data annotation and believe that it will be of great practical value.**

***Index Terms*— medical-imaging, segmentation, segment-anything, foundation-model, finetuning**

I. Introduction

Foundation models are large deep neural networks, often based on the transformer architecture [67], trained on diverse datasets, either with a self-supervised or supervised objective. They learn powerful representations that enable different downstream tasks either through in-context learning or finetuning. They underlay recent advances in language processing [4] and are also gaining importance in computer vision, thanks to the vision transformer (ViT) [13]. The first foundation model that has gained wide-spread adoption for image segmentation is the Segment Anything Model (SAM) [33]. It was trained on a large dataset of natural images with object annotations, using a supervised training objective that mimics interactive annotation. The model supports interactive and automatic segmentation tasks and generalizes to many different imaging modalities. More recently, SAM2 [61] has extended SAM to video data through architectural changes and a large video dataset with objects tracked over time. It supports interactive video segmentation in different modalities.

SAM has been widely studied by the medical imaging community. Initial work has evaluated it for medical segmentation tasks (e.g. [55, 30, 22, 80, 39]). The model showed impressive performance given that it was predominantly trained on natural images, but could not yet compete with domain specific models, especially for difficult tasks such as spine segmentation in MRI [55], segmentation of small organs in CT [26], and other examples [9, 22]. Consequently, follow-up work has improved SAM for medical images, either by finetuning it for interactive segmentation [49, 9, 68, 43] or by using it as a pretrained encoder for semantic segmentation [79, 70, 8, 18, 77, 5]. The model has also been adapted in related domains, for example to improve segmentation in microscopy [1] and histopathology [24, 17]. Some work [49, 9, 18] also tried to build a better *foundation model for medical images*, by finetuning SAM on a large medical dataset and publishing the updated model weights. Some studies have also investigated SAM2 for medical images [12, 83, 50]. They found mixed results for 2D segmentation, improving over SAM for some modalities but with worse performance for others, and promising results for video segmentation [44].

However, a comprehensive study that compares different approaches for *improving SAM as a foundation model for medical images* is so far missing. Specifically, no prior work has explored the impact of different finetuning strategies on the different criteria that a foundation model for medical segmentation should fulfill:

1) It should improve interactive segmentation. The imaging modalities and segmentation tasks in medicine are very diverse. Hence, a single model that can solve any medical segmentation task automatically is currently not feasible[1]. Better interactive segmentation will enable semi-automatic data annotation, leading to faster annotation times, either for data analysis or model training.
2) It should improve the performance for downstream tasks, in particular as a pretrained model for semantic segmentation. This would enable automating segmentation by supervised finetuning, potentially using annotations generated interactively.
3) It should be compatible with the original SAM library tools for data annotation (e.g. [46, 1]), enabling users to benefit from improved interactive segmentation.

Previous work has only addressed one of these aspects at a time: MedSAM [49] and SAMed2D [9] evaluate interactive segmentation (1), Gu et al. [18] study semantic segmentation (2). None of them explicitly study compatibility with user-friendly tools (3).

Institute of Computer Science, University of Göttingen, Germany

[1]There exist some efforts to establish such models for specific modalities, most notably TotalSegmentator [69, 11] for CT and MRI images.

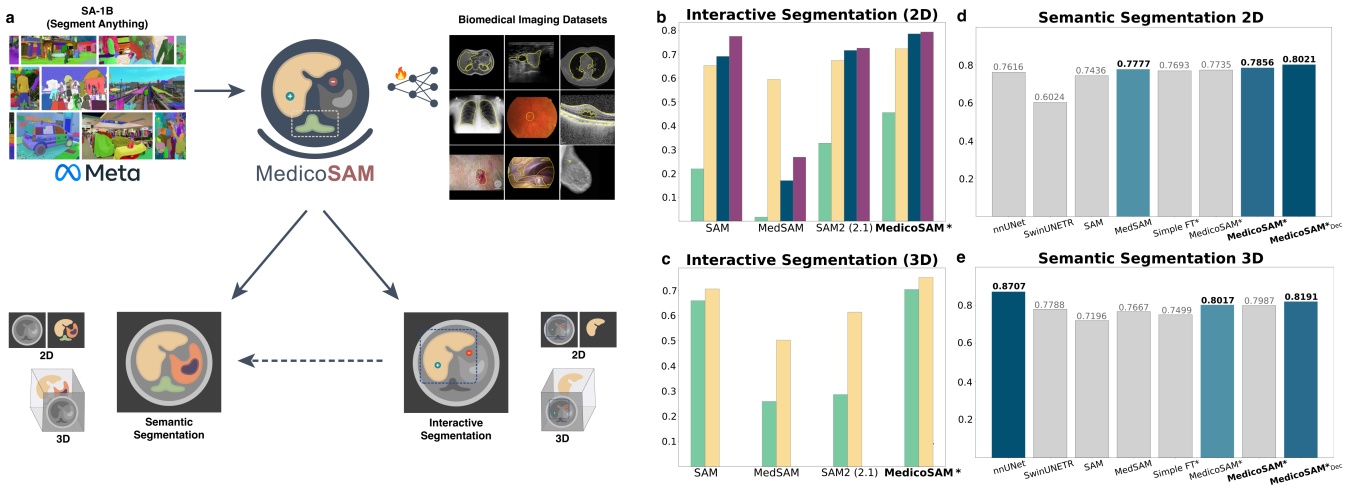

Fig. 1. **a)** Contribution overview: We finetune SAM on a large medical dataset to build MedicoSAM. We evaluate it for interactive and semantic segmentation. The latter requires training on additional annotated data (that could be generated via interactive segmentation), for 2D and 3D data. **b)** Results for interactive 2D segmentation, comparing MedicoSAM and other models derived from SAM. We report the average over 16 datasets for segmentation with a point (green) or box (yellow) prompt and segmentation after iterative correction starting from a point (dark green) or box (dark purple). **c)** Results for interactive 3D segmentation. We report the average over 6 different datasets for segmentation based on a single point or box. **d, e)**. Results for semantic 2D and 3D segmentation. We report the average over 6 datasets in both cases. The three best methods are highlighted in decreasing shades of blue, darker indicates better results, gray otherwise.

Our work closes this gap by comparing different training strategies on the dataset published by SAMed2d [9] and evaluating their effect on (1-3). Specifically, we evaluate own models trained with different strategies and published SAM-derived models on challenging medical segmentation tasks from four different categories: interactive 2D/3D segmentation and semantic 2D/3D segmentation. Where applicable, we also compare to SAM2 and MedSAM2 [50]. An overview of our approach is shown in Fig. 1a and a summary of our results in Fig. 1b-e. We find that domain specific finetuning clearly improves interactive segmentation in 2D and 3D, given the right training objective. For semantic segmentation, pretraining on medical data leads to improved results compared to the original SAM model, with competitive or better performance compared to nnU-Net [28] in 2D but worse performance in 3D. Our software and our best model, which we call MedicoSAM, are available at `https://github.com/computational-cell-analytics/medico-sam`.

II. Methods

We provide a summary of the contributions made by SAM [33], focusing on its training objective (Sec. II-A). We then describe the finetuning strategies explored in our study (Sec. II-B), including our contribution for pre-training semantic segmentation, our extension of SAM to interactive 3D segmentation (Sec. II-C), our methods for 2D and 3D semantic segmentation (Sec. II-D), and our evaluation methodology (Sec. II-E).

A. Segment Anything Model

SAM [33] is a vision foundation model for segmentation tasks. It consists of the image encoder, a ViT [13], the prompt encoder and the mask decoder. This architecture enables the model to solve interactive segmentation tasks based on user input, so called prompts. The image encoder processes the image and outputs an image representation. It contains the majority of parameters. SAM provides three different versions with different encoder sizes, ViT-Huge (ViT-h), ViT-Large (ViT-l) and ViT-Base (ViT-b). The prompt encoder processes the prompts, which can be point coordinates, either a positive point prompt (within the object of interest) or a negative point prompt (outside of the object of interest), a box coordinate, or a low-resolution mask. It outputs a representation of the prompts; point, box and mask prompts can be combined. The mask decoder processes the outputs of image encoder and prompt encoder to predict a mask of the object of interest and a score that estimates the prediction quality. It has two heads. One predicts a single mask and score, the other predicts three masks and scores. The second head is for the case of a single point prompt, which can result in ambiguities for part-object segmentation. Fig. 2a shows an overview of SAM's architecture with additions made by us marked in orange.

SAM was evaluated for a wide range of segmentation tasks in diverse image modalities, showing remarkable generalization. These capabilities are mainly due to two factors: its large and diverse training set and sophisticated training objective. The training dataset, called SA-1B, consists of 11 million images with 1 billion annotated objects. It was generated by human annotators using SAM for semi-automatic annotation, followed by retraining and further annotation with the updated model, repeated multiple times. The model was trained on this dataset using a supervised training objective that mimics interactive object annotation and correction: For a given ground-truth mask, the objective first samples either a point or box prompt and then corrects the model predictions with point prompts in multiple steps. In each step it computes L_{mask}, the loss between true and predicted mask as well as L_{iou}, the loss between the intersection over union (IOU) of true and predicted mask and the predicted score. These

losses are accumulated and averaged at the end of a training iteration. See Alg. 1 for the pseudo-code of a training iteration. Besides this objective, the training proceeds as usual for deep neural networks by updating model weights with a version of stochastic gradient descent over multiple epochs.

Algorithm 1: The training objective of SAM [33], according to [1]. We have added the optional joint pretraining of the segmentation decoder ($e_{sem} = 1$).

Input: Images and target masks, hyperparameters n_{obj}, n_{steps}, p_{box}, p_{mask}, e_{sem}
Output: Updated model parameters

1 Sample minibatch of images and target masks
2 Sample fixed number of object masks n_{obj} per image
3 Predict embeddings for the images with the encoder
4 Initialize empty list L for losses
5 **for** mask m in minibatch **do**
6 Initialize empty list for prompts p
7 Sample u_{box} uniformly from $[0, 1]$
8 **if** $u_{box} < p_{box}$ **then**
 `// The box can also be distorted`
9 Compute bounding box of m, add as prompt to p
10 **else**
11 Sample random point from m, add as positive point prompt to p
12 Apply prompt encoder to p
13 **if** p contains single point prompt **then**
14 Predict masks and IOUs with multi mask head of mask decoder
15 Select predicted mask $\hat{m}$ and IOU $\hat{i}$ with the highest IOU value
16 **else**
17 Predict mask $\hat{m}$ and IOU value $\hat{i}$ with single mask head of mask decoder
18 Compute mask loss $L_{mask}(\hat{m}, m)$
19 Compute IOU i between $\hat{m}$ and m
20 Compute regression loss $L_{iou}(\hat{i}, i)$
21 Add L_{mask} and L_{iou} to L
22 **for** j = 1 **to** n_{steps} **do**
23 Sample positive point from $m \,\&\, !\hat{m}$, add to p
24 Sample negative point from $!m \,\&\, \hat{m}$, add to p
25 Sample u_{mask} uniformly from $[0, 1]$
26 Remove mask prompt from p if present
27 **if** $u_{mask} < p_{mask}$ **then**
28 Add $\hat{m}$ to p
29 Run lines 12-21 with current p
30 **if** e_{sem} **then**
31 Compute binary target b as union of all target masks
32 Predict binary mask $\hat{b}$ with additional segmentation decoder
33 Compute $L_{mask}(\hat{b}, b)$ and add it to L
34 Average losses in L, perform backprop
35 Update model parameters via optimizer

Recently, SAM2 [61] has adapted this approach to interactive video segmentation by adding a memory bank that stores prompts and mask predictions from previous frames. SAM2 was trained on a large annotated video dataset. While SAM2 is promising in the medical domain to analyze videos or volumetric data, it has so far received fewer attention. Some studies [50, 12, 83] have evaluated it and, to our knowledge, only MedSAM2 [50] has provided an improved version of the model. Here, we include SAM2 and MedSAM2 in the evaluation for interactive segmentation, but do not finetune the model on medical data due to its recency and missing support in tools for medical data annotation.

B. Finetuning Segment Anything

We compare, implement, and extend different approaches for improving SAM for medical images by finetuning on a labeled dataset of medical images using a variation of Alg. 1. This training algorithm has not been published by SAM [33], so each publication has used a custom implementation. The finetuning strategies also differ in which model parts they update, and whether an adapter-based strategy like LoRA [25] is used.

The authors of MedSAM [49] assemble a dataset of 1.5 million masks from CT, Endoscopy, MRI, X-Ray and other modalities, based on published data. They introduce a simple training objective that uses only box prompts, which are derived during training from the mask annotations. This objective corresponds to $n_{steps} = 0$, $p_{box} = 1$, $p_{mask} = 0$ in Alg. 1. They use the ViT-b encoder and update all parameters of the image encoder and mask decoder, freezing the prompt encoder. The finetuned model is evaluated for interactive segmentation based on box prompts. The model is publicly available.

The authors of SAM-Med2D [9] build the SA-Med2D-20M dataset [73], which consists of 20 million masks in 5 million images. The data covers ten different modalities (CT, Endoscopy, MRI, ultrasound, X-Ray and others) and is collected from published data. They finetune using with a training objective of $n_{steps} = 8$, $p_{box} = 0.5$, $p_{mask} = 1$. They deviate from Alg. 1 in two ways: they sample either 1, 3, 5 or 9 points per step instead of a single positive and negative one (cf. lines 23-24) and they do not compute gradients for the prompt encoder in the steps corresponding to lines 22-29. They use ViT-b as image encoder and insert a low rank projection layer between attention and feed-forward layer of each transformer block, similar to [70]. Only these parameters of the image encoder are updated, others frozen. For prompt encoder and mask decoder all parameters are updated. In addition, they train the model with a smaller input image size of 256 x 256 pixels instead of 1024 x 1024 pixels used by SAM. They also study a simpler finetuning strategy called FT-SAM where only the mask decoder is updated during training, image and prompt encoder are frozen. The finetuned models are evaluated for interactive segmentation and are publicly available.

The authors of [18] assemble a dataset comprising 100,000 masks in 300,000 images from CT, MRI, X-Ray and ultrasound. This dataset also contains unlabeled images used for self-supervised training. They compare two different finetuning objectives: updating the image encoder in a self-supervised manner using MAE [21] and using a simple supervised strategy with a box or a single point prompt, corresponding to $n_{steps} = 0$, $p_{box} = 0.5$, $p_{mask} = 0$. They further compare the different model sizes (ViT-b, ViT-l, Vit-h) with and without the use of LoRA [25]. They evaluate these models for semantic segmentation tasks. None of the models are publicly available.

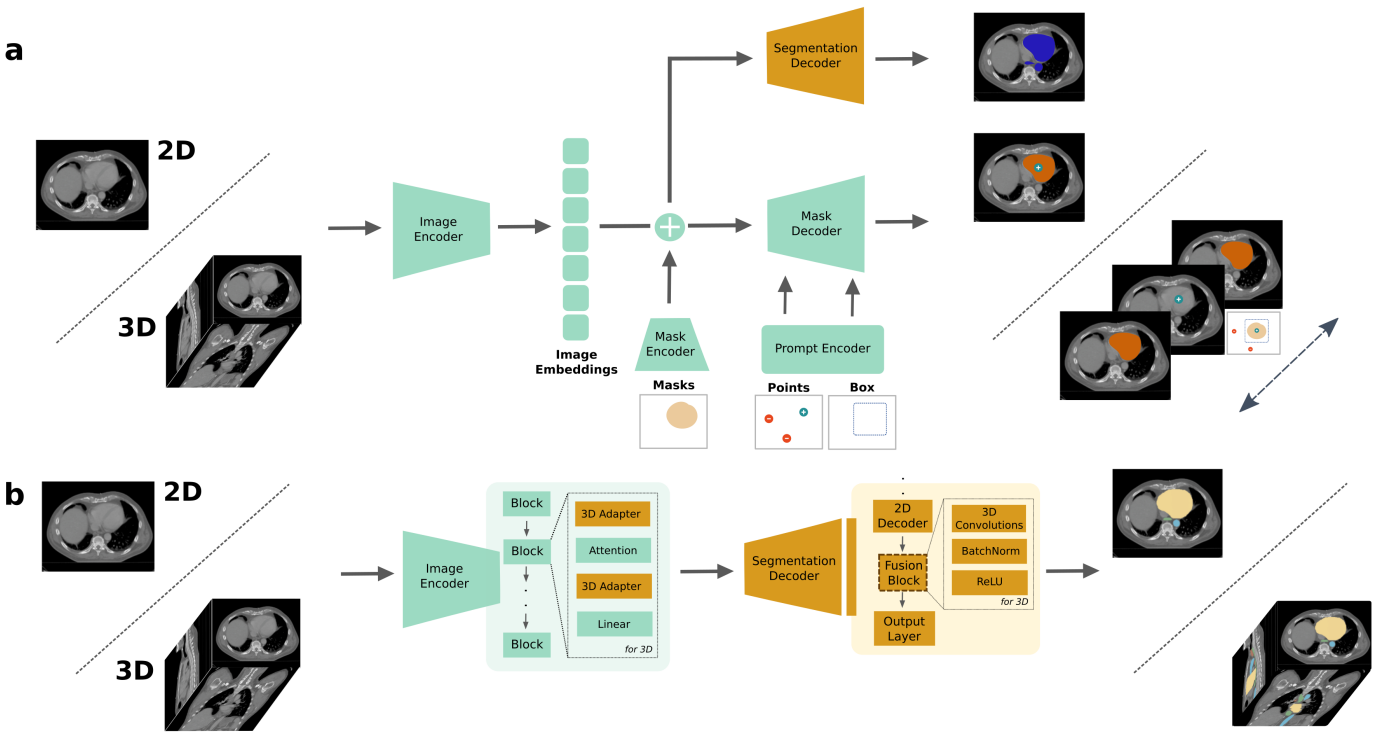

Fig. 2. **a)** The SAM architecture for interactive segmentation consists of image encoder, prompt encoder (split into a part for mask prompts and for point/box prompts), and mask decoder. For 3D interactive segmentation, we propagate prompts across the depth axes. In addition, we add a convolutional decoder for automated segmentation (orange). This decoder is pre-trained with a binary segmentation task (blue masks), jointly with training for interactive segmentation.
b) Adaptation for 2D/3D semantic segmentation. A convolutional decoder predicts the segmentation output (same as "Segmentation Decoder" in a)). For 3D segmentation, additional adapters are added to the image encoder and outputs of the segmentation decoder are computed per slice, stacked, and processed by a 3D convolution. Architecture modifications are highlighted in orange.

In summary, prior work has studied different finetuning objectives and different model update strategies. To analyze the influence of the objectives we build on the versatile implementation of Alg. 1 provided by μSAM [1], which was developed for microscopy data. We extend this implementation to also support simpler schemes (e.g. $n_{steps} = 0$) within the same framework. We do not study self-supervised training and we finetune all model parts, without the use of adapter layers. The first choice is due to the fact that self-supervised training would very likely lead to a loss of interactive segmentation performance. The second choice because we want to provide models compatible with the SAM library and tools using it. Introducing adapter layers would make the model incompatible and thus not practically useful, see also Sec. III-C. Consequently, we study three different finetuning strategies:

- MedSAM (adapted from [49]) uses only a box prompt, corresponding to $n_{steps} = 0$, $p_{box} = 1$, $p_{mask} = 0$ in Alg. 1.
- SimpleFT (adapted from [18]) uses a single box or a single point prompt, corresponding to $n_{steps} = 0$, $p_{box} = 0.5$, $p_{mask} = 0$ in Alg. 1.
- MedicoSAM uses the full objective with $n_{steps} = 8$, $p_{box} = 0.5$, $p_{mask} = 0.5$[2].

[2]We have trained two different versions of this model, one with $p_{mask} = 0.5$ and one with $p_{mask} = 0$, see for details.

We set $n_{obj} = 5$, use the Dice loss for L_{mask} and the L2 loss for L_{iou} in all cases. We also add an option to jointly pre-train the decoder for semantic segmentation. In this case, the decoder predicts a binary mask and we compute a loss between this prediction and the union of all target masks, corresponding to $e_{sem} = 1$ in Alg. 1. See Sec. II-D for details on the decoder architecture. We finetune the models on SA-Med2D-20M [73]. We also benchmark the published models MedSAM [49], SAM-Med2D [9], FT-SAM [9], SAM [33], and where applicable SAM2 [61] and MedSAM2 [50]. We use ViT-b for all models, expect for MedSAM2, where only ViT-Tiny (ViT-t) is available.

C. Interactive 3D Segmentation

Unlike SAM2, which supports image and video segmentation, SAM can only segment 2D images. Follow-up work has implemented interactive segmentation for videos or volumetric data in medical images, e.g. for CT [41, 52]. Here, we use the implementation from [1], which is based on prompt propagation. Briefly, SAM segments an object in one or multiple slices based on given prompts. Then, the segmentation mask(s) are projected to adjacent slices, prompts are derived from them, and segmentation is run for these prompts. The process is repeated until the object is segmented throughout the whole volume or a stopping criterion based on the IOU between

adjacent slices is met. Multiple options are provided for deriving prompts from projected masks: using a single positive point prompt placed at the mask's center, using multiple point prompts derive from the mask, using the bounding box derived from the mask, using the bounding box and low-resolution version of the mask, and combinations of these options. A user can correct the segmentation by annotating slices with manual prompts and rerunning the segmentation.

D. Semantic Segmentation

SAM itself can only be used for interactive segmentation. In [33] a method for automatic instance segmentation, called automatic mask generation, is proposed. However, it does not support semantic segmentation, which is more relevant for medical images. Conceptually, SAM does not learn explicit semantic knowledge, since it is only trained to distinguish individual objects in an image. To evaluate different pretrained SAM models for semantic segmentation, we further finetune them for specific tasks using annotated data, based on architectural changes for 2D and 3D segmentation.

For 2D segmentation we add a UNETR-like [20] convolutional segmentation decoder. It processes and upsamples the embeddings predicted by the image encoder, in this case the pretrained ViT-b encoder, to predict a semantic segmentation. Prompt encoder and original mask encoder of SAM are discarded. This model is trained for semantic segmentation with a loss between predictions and semantic labels. Here, we use a combination of Cross Entropy and Dice loss. We also study a variant where the segmentation decoder is pretrained, see Sec. II-B for details. The segmentation decoder is marked in orange in Fig. 2a. The choice of this architecture for semantic segmentation was motivated by similar approaches for automatic segmentation with SAM in [1, 71].

For 3D segmentation we adopt the implementation of MA-SAM [8] to extend the image encoder to volumetric data. This is achieved by flattening the batch and depth dimensions so that the image patches extracted from an input volume can all be processed by the encoder. To make use of depth information, additional adapter layers are introduced. These layers decrease the number of features per token, rearrange tokens into a volumetric representation, apply a 3x1x1 convolution, project the number of features back, and flatten the batch and depth axis. Each transformer layer is augmented with two of these adapters, one before and one after the attention layer. The parameters of the adapters are randomly initialized. Unlike MA-SAM, which re-uses SAM's mask decoder, we use our new 2D segmentation decoder. We apply it slice-by-slice to the image embeddings, then stack its outputs and apply a 3D convolution layer to predict a volumetric segmentation. This design enables to also initialize the pretrained decoder weights. The architecture is shown in Fig. 2b.

E. Evaluation

We compare models for interactive and semantic segmentation. For semantic segmentation (2D and 3D), we follow standard procedures and compare the predicted semantic masks with ground-truth annotations using the Dice coefficient.

For interactive 2D segmentation we adopt the evaluation procedure of [1]. This approach simulates iterative user-based annotation. It requires object mask annotations. For a given object, a single prompt is sampled, either a point or a box. The object is then iteratively corrected by sampling point prompts from errors in the prediction. In each iteration a positive point prompt is sampled from the region where the prediction is missing (prediction is negative, annotation is positive) and a negative point prompt is sampled from the region where the prediction should not be (prediction is positive, annotation is negative). The Dice coefficient between true mask and prediction is computed for the initial segmentation and each correction iteration. For interactive 3D segmentation we evaluate the initial segmentation derived from a point prompt (randomly sampled from the object in the central slice) and from a box prompt (also in the central slice). We do not simulate iterative correction of the masks due to the higher computational demand of 3D segmentation. We run a grid search over the different options for deriving prompts, see Sec. II-C, on separate validation data.

III. Experiments

We finetune different models based on SAM with ViT-b encoder, initialized with the weights from [33]: MedSAM, SimpleFT and MedicoSAM, see Sec. II-B for the respective training strategies. The models are trained on the publicly available subset[3] of SA-Med2D-20M [73], which contains 3.7 million images and 15.8 million masks. To compare the impact of p_{box} and n_{steps} (see Alg. 1), we train a model for each of the three configurations on 60% of the data (50% train, 10% val), using a stratified split over different modalities, setting $p_{mask} = 0$, $e_{sem} = 0$. We train another version of MedicoSAM on the full dataset (90% train, 10% val; same val split as before), with $p_{mask} = 0.5$ and $e_{sem} = 1$. The latter model is included to study the impact of masking and semantic decoder pretraining. The smaller subset for the three initial models is chosen to reduce training cost. The models are trained on 8 A100 GPUs with 80GB VRAM for 300,000 iterations with a batch size of 7 per GPU, corresponding to 21 epochs for models trained with 60% data, and 11 epochs for the model trained on the entire data. We use the AdamW optimizer [47] with an initial learning rate of $1e-5$ and a scheduler that reduces the learning rate by a factor of 0.9 after each epoch. We evaluate these four models, the original SAM and SAM2 models, and published derived models for interactive segmentation (2D and 3D), semantic segmentation (2D and 3D), and integration with user-friendly tools.

A. Interactive Segmentation

We evaluate ten different models for interactive 2D segmentation. We use 16 datasets that are not part of our training dataset to evaluate generalization capabilities. These datasets represent a variety of medical segmentation tasks from CT [57, 3, 58], dermoscopy [14], endoscopy [64, 53], MRI [51, 16], ophthalmology [35], ultrasound [38, 48, 32], and X-Ray [29, 40, 7, 78]. Here, 3D dataset are split into separate images.

[3]The dataset contains a private test set of 920k images and 4 million masks.

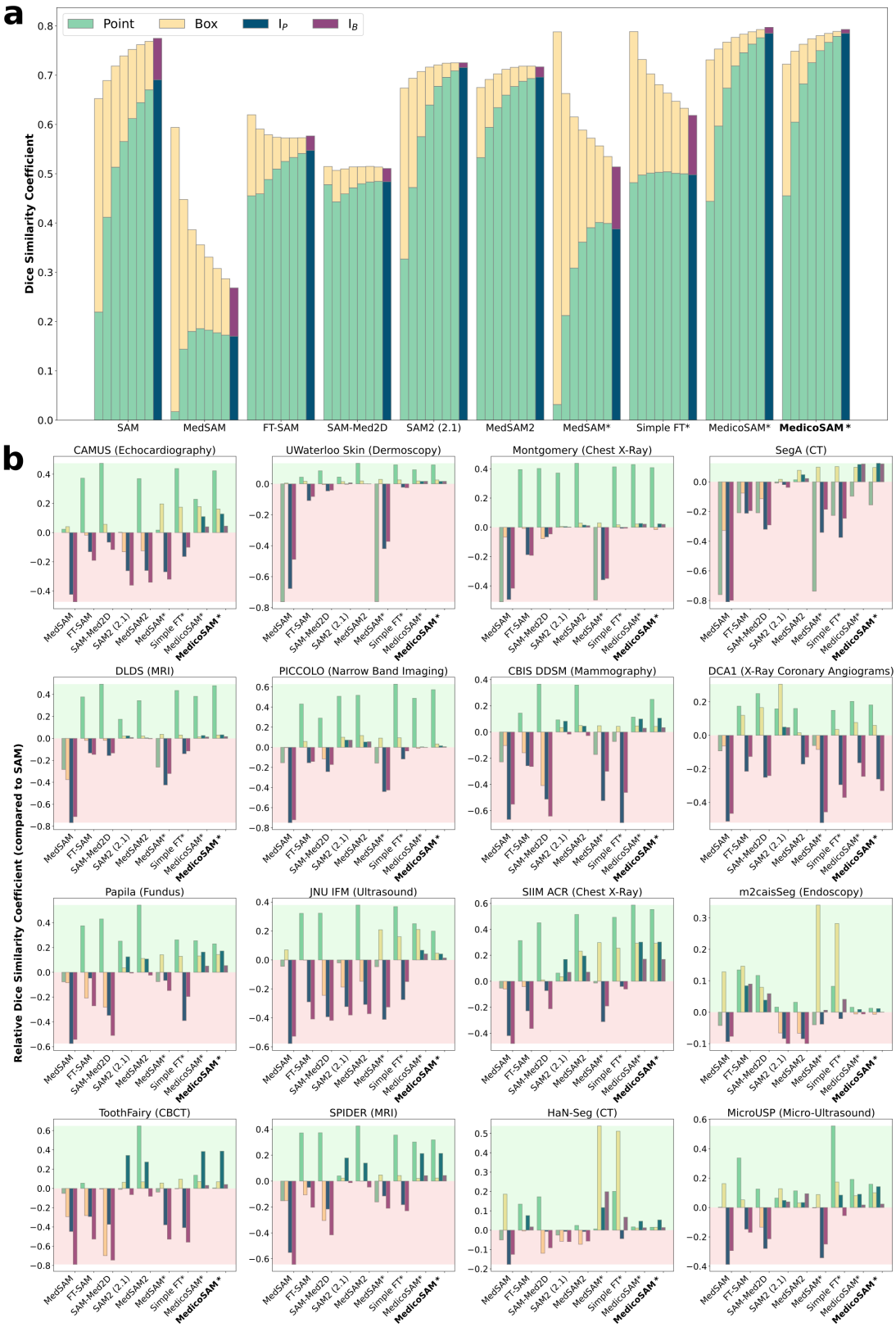

Fig. 3. **a)** Overall results for interactive 2D segmentation. We report the Dice coefficient for simulated interactive segmentation. Each bar corresponds to the result of a correction iteration, starting either from a point (green) or a box (yellow) prompt. The result after correction is highlighted in dark green / dark purple. We compare 10 different models. Models trained by us are marked with a * and the model trained on the entire dataset is marked in bold font. The same model notation is used in all figures. **b)** Interactive segmentation results for 16 individual datasets. We report the absolute difference of the Dice coefficient compared to the original SAM and report only the results for the initial and final segmentation.

The results averaged over all datasets are shown in Fig. 3 a). We report the results for initial prompt-based segmentation and correction iterations, see also Sec. II-E, using a point or box as initial prompt. Fig. 3 b) shows the results for individual datasets, where we only report the results for the initial and final segmentation, corresponding to the last correction iteration. We report the difference in the Dice coefficient compared to SAM. MedicoSAM is the only model that clearly improves upon SAM in all settings. The other finetuned models either only improve for the segmentation with a single prompt (MedSAM*, SimpleFT*, MedSAM2) or lead to an overall worse segmentation. Models trained with $n_{steps} = 0$ yield worse results for an increasing number of prompts. SAM2 performs slightly worse compared to (original) SAM.

A qualitative comparison of results and image embeddings for different models is shown in Fig. 5. We also study the statistical reliability based on five random seeds on a single dataset in Fig. 6a, where we observe that the standard deviation of results is low compared to differences in the model performance. We perform an additional study to further understand the effect of n_{steps} and p_{mask} in Fig. 7.

We evaluate interactive 3D segmentation with a single point or box derived from segmentation annotations for 6 different external datasets from MRI [56, 51], CT [23, 65], and ultrasound [32, 36]. We compare six different models. For the models based on SAM, we use the method described in Sec. II-C and we find the best setting with a grid search on a separate validation set. SAM2 supports interactive 3D segmentation as is (by interpreting the 3D data as a video), and we do not perform a grid search to optimize parameters for inference (same for MedSAM2). The overall results are shown in Fig. 1 c) (except for SimpleFT) with results for individual datasets in Fig. 4. Only MedicoSAM improves consistently. The fact that SAM2 and MedSAM2 are used as is, without optimizing parameters in a grid search, may disfavor them.

B. Semantic Segmentation

We evaluate different pretrained SAM models for semantic segmentation in 2D and 3D, using the implementations described in Sec. II-D. For 2D segmentation we use 6 external datasets from dermoscopy [10], mammography [40], narrow band imaging [64], optical coherence tomography [74], panoramic radiographs [63], and X-Ray angiography [7]. For 3D segmentation we use 6 external datasets from CT [65], MRI [51, 54, 56], and ultrasound [32, 36]. We use separate splits for training and evaluation. The results are shown in Fig. 8. We also report the results for nnU-Net [28] and Swin UNETR [19] trained on the same splits, using the default nnU-Net v2 setup and the MONAI [6] implementation, respectively. We also report the results for BiomedParse [81], a foundation model for text-based segmentation trained on a large biomedical imaging dataset. We apply it to the 2D datasets for which we could find a modality prompt matching the training data of BiomedParse.

Here, we see an advantage in using specifically pretrained backbones (MedSAM, SimpleFT, MedicoSAM) compared to the initial SAM model. We also see an advantage in using the pretrained segmentation decoder. These findings are consistent across 2D and 3D datasets. In 2D, SAM-derived models are competitive or slightly better than nnU-Net and better than Swin UNETR. In 3D, nnU-Net is the best method, in particular due to the poor performance of SAM-based models on two of the datasets. Note that nnU-Net and Swin UNETR are trained from scratch, whereas SAM-derived models are initialized with a pretrained encoder and in one case also a pretrained decoder. BiomedParse is applied to without further training and where applicable performs at-par with the best MedicoSAM model.

C. Tool Integration

An important practical aspect in improving SAM as a foundation model is the integration with tools for data annotation. Hence, models should not introduce changes to the architecture that make it incompatible with the SAM library and tools using it. We check this for three models, MedSAM, SAM-Med2D and MedicoSAM, and four different tools: two napari plugins [15, 1] and two 3D Slicer extensions [46, 75]. Tab. I shows the compatibility. We find that SAM-Med2D does not work in any of the tools because it uses adapters and changes the image input size. MedSAM and MedicoSAM work in all of the tools with at most small code changes. We qualitatively compare the models for data annotation with these tools in Fig. 9.

Tool / Model	MedSAM	SAM-Med2D	MedicoSAM
SegmentWithSAM	✓	×	✓*
SAMM	✓	×	✓*
napari-sam	✓	×	✓*
μSAM	✓	×	✓

TABLE I
COMPATIBILITY WITH USER-FRIENDLY TOOLS, FOR THREE MODELS AND FOUR GRAPHICAL TOOLS SUPPORTING SAM FOR DATA ANNOTATION. THE * REPRESENTS MINOR CODE CHANGES NECESSARY TO ADAPT A FILE PATH OR URL TO LOAD DIFFERENT WEIGHTS.

IV. Conclusion and Discussion

We have comprehensively studied how to improve SAM [33] as a foundation model for medical images, by evaluating the impact of different finetuning objectives on interactive and semantic segmentation. We found that interactive segmentation improved clearly, compared to original SAM and SAM2 [61], but critically dependent on the choice of objective. For semantic segmentation, we found that domain specific pretraining also provides a benefit. However, the segmentation quality is only modestly better than nnU-Net [28] in 2D and worse in 3D; despite nnU-Net not being pretrained. We also argued that models based on SAM should adhere to the original architecture to enable integration with user-friendly tools. This is especially important due to their improvement in interactive segmentation, which relies on such tools for practical value. We have published our best model, MedicoSAM, and believe that it will be of great practical value for data annotation. Furthermore, we believe that our findings will also prove

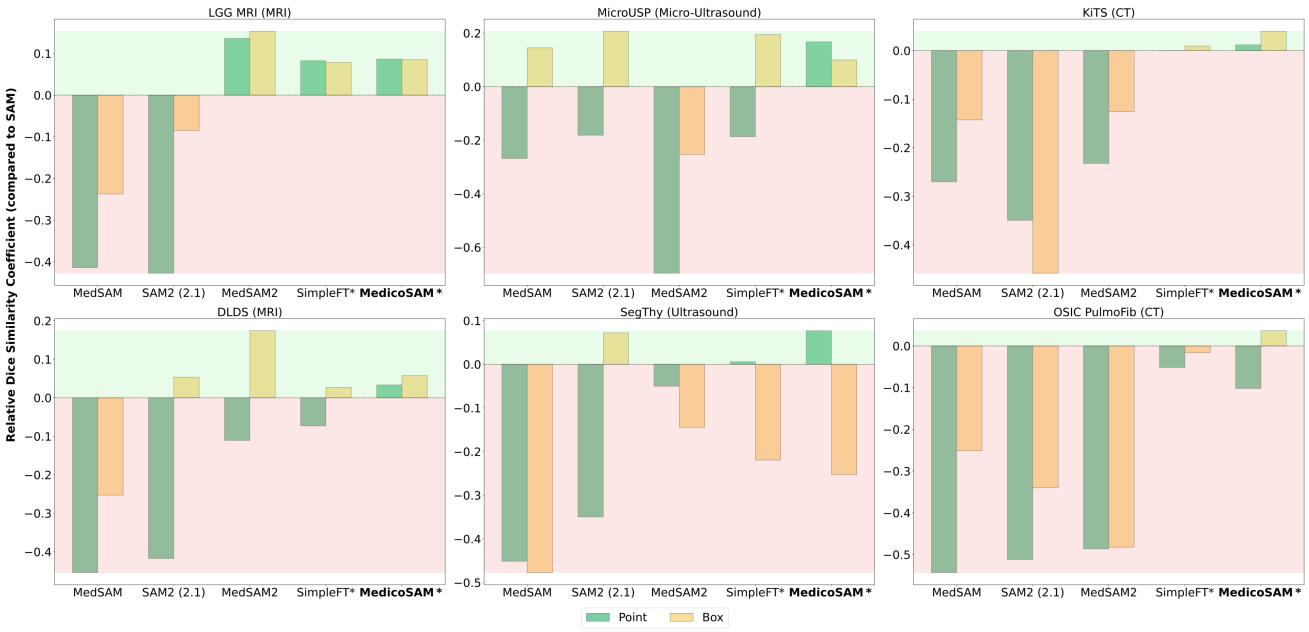

Fig. 4. Results for interactive 3D segmentation for 6 different datasets. We report the difference in Dice score compared to SAM for four other models. Segmentations are derived from a single point (green) or box (yellow) prompt placed in the central slice for each object in the respective dataset. We use the implementation of [1] for methods using SAM, determining the best method for prompt propagation on a separate validation set, see also Sec. II-C. SAM2 supports 3D segmentation by default.

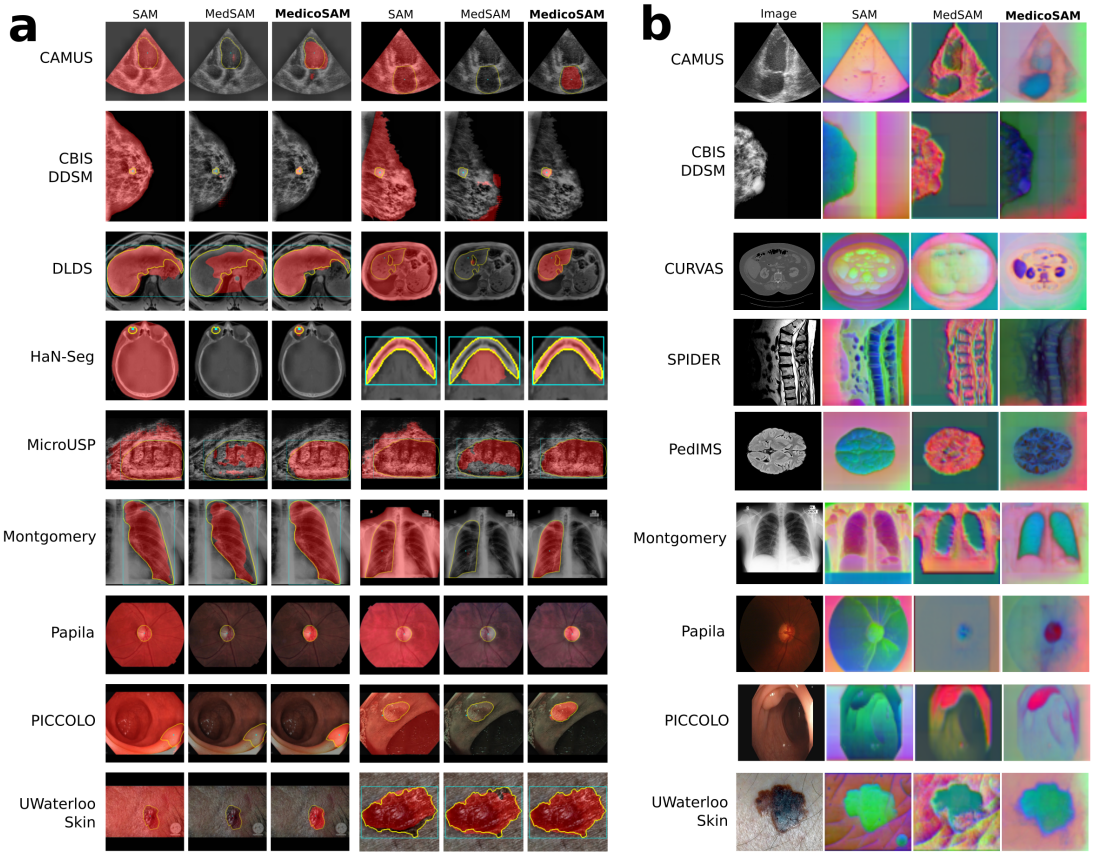

Fig. 5. **a)** Qualitative results for interactive 2D segmentation. We compare interactive segmentation based on a single point or single box prompt (cyan) with SAM, MedSAM and MedicoSAM for nine different datasets. For each image, we show prompts with a large improvement of MedicoSAM over SAM and the corresponding MedSAM result. **b)** Outputs of the image encoder from the three different models on different datasets, additionally Abdominal CT [62] and Brain MRI [59] visualized by their three main PCA components. MedSAM and MedicoSAM seem to learn a more discriminative representation with clearer distinction of background.

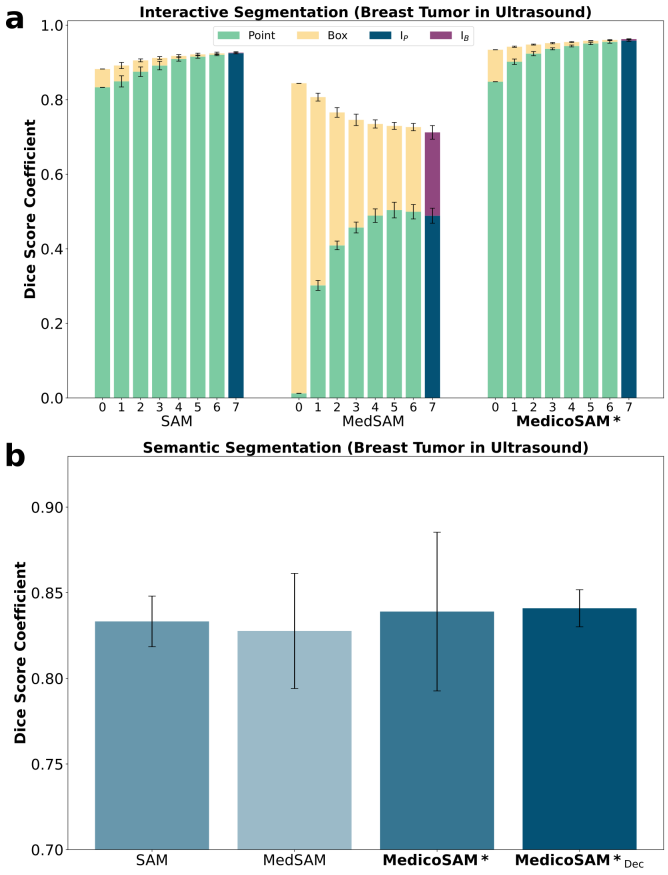

Fig. 6. **a)** Statistical analysis of interactive segmentation on the ABUS data for breast tumor segmentation in ultrasound images [2] with three models. We run each interactive segmentation experiment five times with different random seeds and report standard deviations as error bars. The deviations are small compared to differences in model performance. **b)** Statistical analysis of semantic segmentation (same data as a)). We run the training for each of the four models five times with different random seeds. The deviations are larger than performance differences, the model with pretrained decoder shows a lower deviation.) The methods are colored in decreasing shades of blue, darker indicates better results.

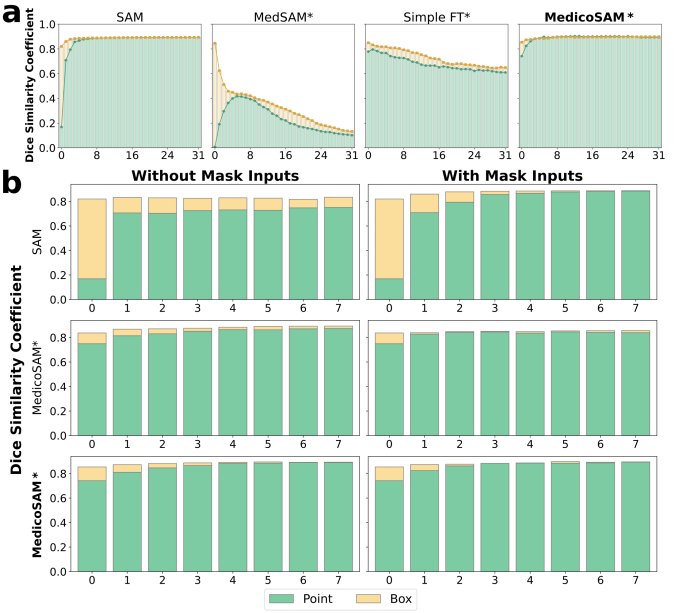

Fig. 7. **a)** Ablation study for the influence of $\boldsymbol{n_{steps}}$ in Alg. 1. We run interactive segmentation with 32 correction iterations, two with $\boldsymbol{n_{steps}} = \mathbf{0}$ (MedSAM, SimpleFT), two with $\boldsymbol{n_{steps}} = \mathbf{8}$. The former show a degrading performance with further prompts, the latter show increasing performance, but plateau after 6-8 iterations. **b)** Ablation study for the influence of $\boldsymbol{p_{mask}}$, comparing models trained with $\boldsymbol{p_{mask}}$ = 1, 0, 0.5 (SAM, MedicoSAM*, **MedicoSAM***) and interactive segmentation without and with use of the previous mask prediction as prompt for the next iteration. SAM and MedicoSAM* perform slightly better in the setting corresponding to their training, **MedicoSAM*** is robust.

valuable to adapt recent and future foundation models, e.g. SAM2, to medical imaging.

Our main goal was the comparison of different objectives for finetuning SAM on large medical data. We found that it is crucial to use an objective that trains segmentation with box and mask prompts for multiple training iterations, i.e. $0 < p_{box} < 1$ and $n_{steps} > 0$ (Alg. 1). Otherwise we observed a catastrophic forgetting-like effect, where the model did not yield accurate results for interactive segmentation with a point prompt or with multiple prompts, see results for MedSAM and FT-SAM in Fig. 3. This observation is especially important since MedSAM, the most cited SAM version for medical images, performs worse than the original SAM in many settings due to its choice of objective, which has been independently reproduced, e.g. by MedSAMix [72]. While not explicitly studying other values than $p_{box} = \{0.5, 1\}$, we assume that values smaller than 0.5 would favor the performance with point prompts and vice versa, but would likely not yield to catastrophic forgetting. A value of 0 would likely result in a model that yields poor results in response to box prompts. We found that models trained with a value of $n_{steps} = 8$ were robust to segmentation with a larger number of prompts, i.e. their performance did not decrease with increasing number of prompts as is the case for $n_{steps} = 0$. However, their performance plateaus. See Fig. 7a. Training with a larger n_{steps} value could potentially delay the onset of this plateau and increase its height, i.e. lead to better segmentation with more prompts. Finally, we also studied the effect of p_{mask}, where we found that models trained with $p_{mask} = 0$ perform better without a mask prompt in interactive segmentation, models trained with $p_{mask} = 1$ perform better with a mask prompt, and models trained with $p_{mask} = 0.5$ are robust. See Fig. 7b. While this trend is expected, interestingly, the difference in performance is not very pronounced, i.e. a model trained with $p_{mask} = 1$ does not learn to rely on the presence of a mask prompt and vice versa.

We also studied the effect on semantic segmentation by finetuning adapted architectures for specific segmentation tasks. Here, we found that SAM-derived models performed modestly better than SAM, but that the exact objective used for pretraining them on the large medical dataset did not have a big impact (similar results for MedSAM, MedicoSAM and SimpleFT). A further modest improvement could be gained by using a pretrained segmentation decoder. See results in Fig. 1 d,e and Fig. 4. Note that the best SAM-derived model performs only modestly better than nnU-Net [28] in 2D segmentation and worse than in 3D, despite the fact that nnU-Net was not pretrained. This highlights that the advantages

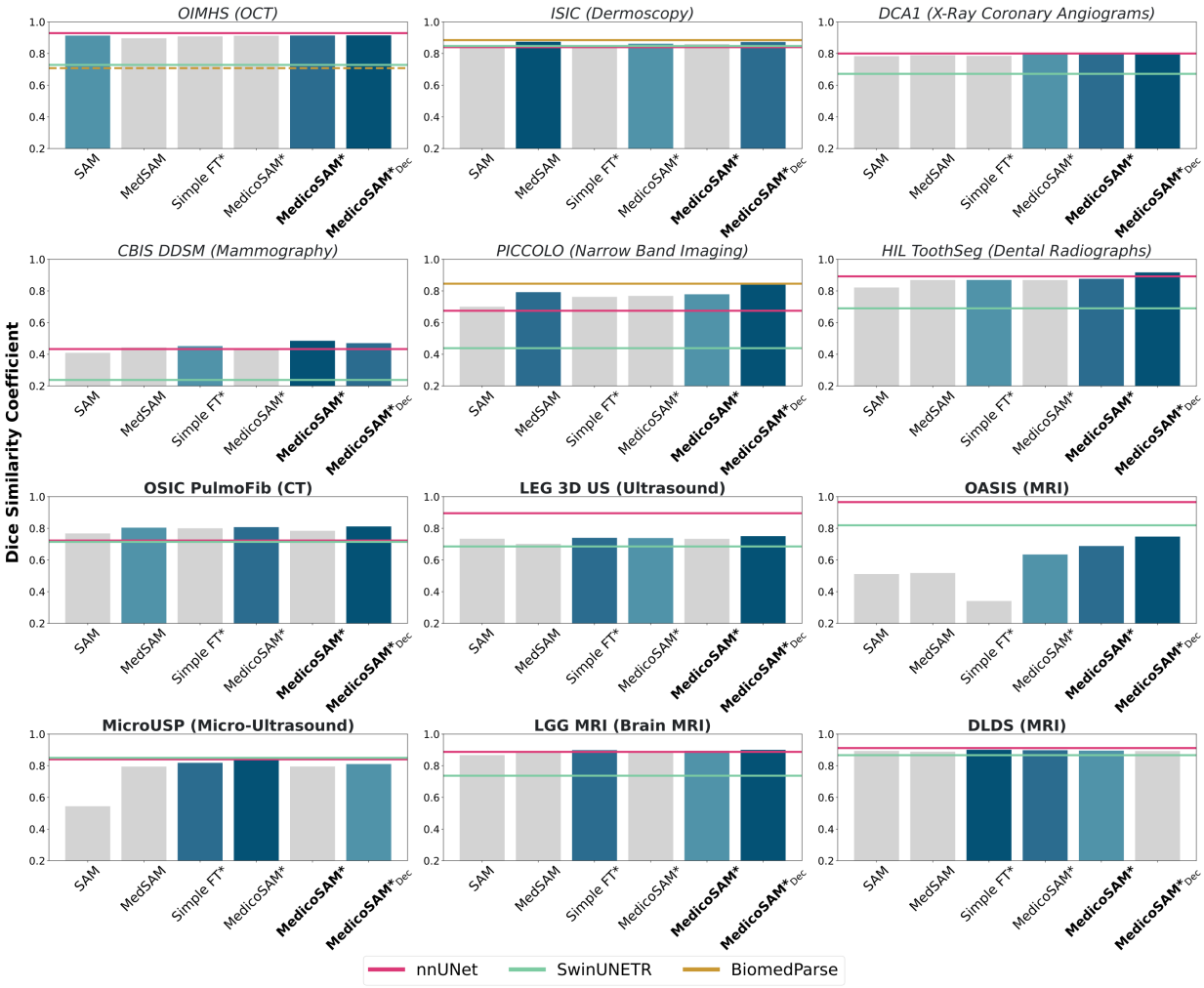

Fig. 8. Semantic segmentation results for 2D and 3D segmentation. The names in italic font indicate 2D data and in bold font 3D data. We compare different pretrained models, original SAM, MedSAM, SimpleFT and MedicoSAM in three versions: trained on the reduced training set, trained on the entire training set (bold font) and trained on the entire training set and using decoder initialization (bold font, subscript "Dec"). The three best SAM-derived methods are colored in decreasing shades of blue, darker indicates better results, gray otherwise. Results for reference methods are indicated by horizontal lines. The dotted horizontal line indicates a case where only a subset of classes could be segmented with BiomedParse.

of SAM-derived models for semantic segmentation in medical images are not yet clear, at least with our implementation. In this context, the finding of [27] that transformer-based segmentation architectures generally do not perform better than nnU-Net is also relevant. Further improvements would likely require updates of the adaptation strategy, especially an extension to 3D pretraining for volumetric segmentation.

Another important emergent trend are foundation models that can segment images based on text prompts. They are especially promising for semantic segmentation without the need for further training. We have included a comparison to one such model, BiomedParse [81] in selected experiments. There are further models, in many cases based on the CLIP [60] architecture, for medical imaging [82, 42, 45]. MedCLIP-SAM [34] combines CLIP and SAM for text based segmentation. Hence, a promising future avenue of research is the combination of such a model with our findings, to obtain a foundation model that can address both SAM-style interactive and text-based segmentation.

Other related research, e.g SAM-REF [76], has investigated the effect of late and early fusion of image features and prompts. SAM uses a late fusion approach. Early fusion can lead to better segmentation of fine-grained structures, while being less efficient because it couples image and prompt processing. Hybrid approaches can provide a good trade-off. This work could profit from our findings on finetuning objectives, as these are orthogonal to the feature alignment mechanism. Furthermore, recent work on parameter efficient finetuning of SAM for biomedical images has found that strategies like LoRA do not provide better segmentation quality, but can lead to more efficiency [66]. These findings likely directly apply to semantic segmentation finetuning in our case.

Acknowledgment

The work of Anwai Archit was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - PA 4341/2-1. The work of Luca Freckmann was funded by the DFG under Germany's Excellence Strategy - EXC 2067/1-390729940. This work is supported by the Ministry of Science and Culture of Lower Saxony through funds from the program

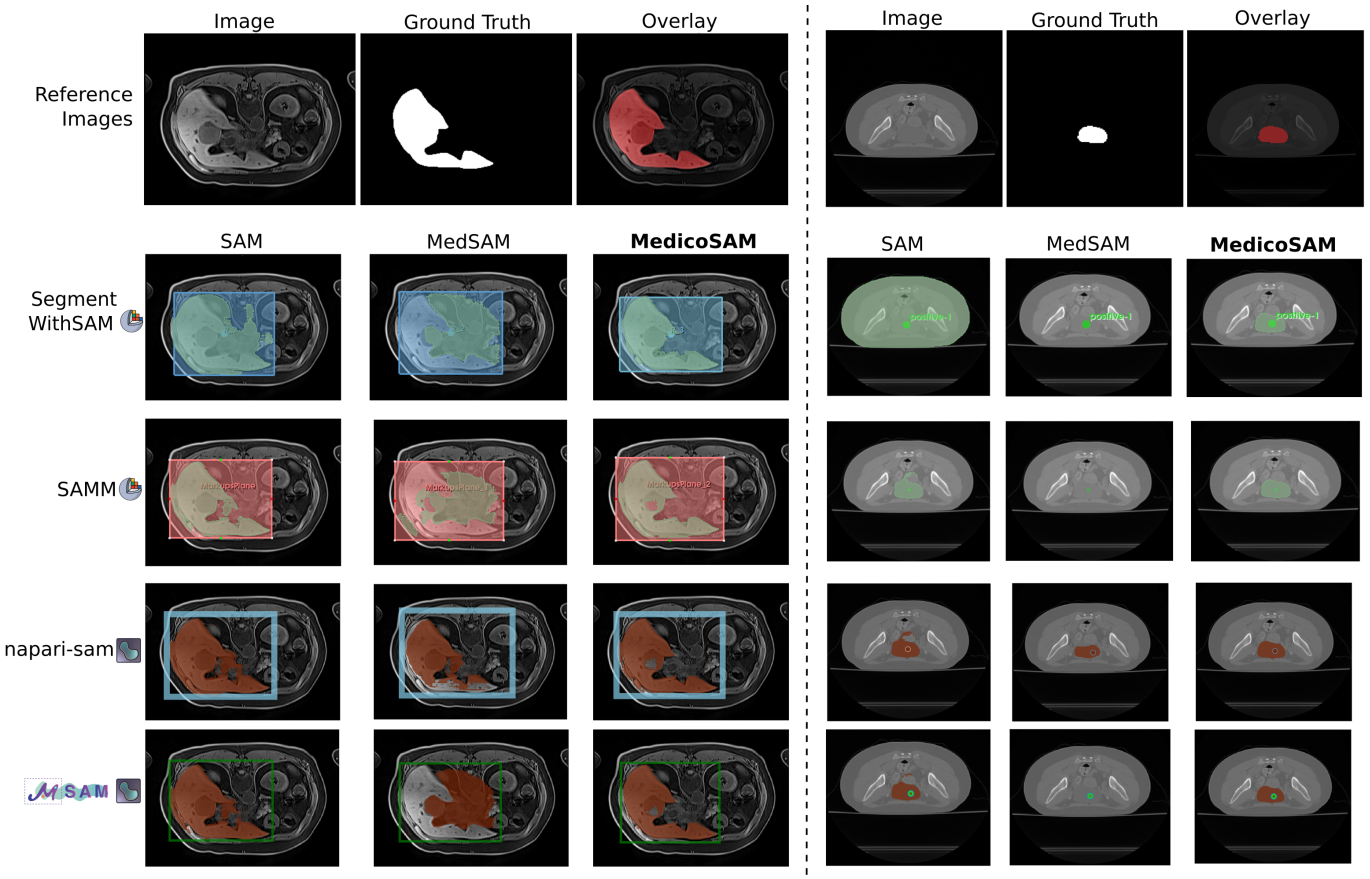

Fig. 9. Using SAM, MedSAM and MedicoSAM in four different tools for interactive segmentation. The top row shows the reference image and the object that is annotated. The four rows below show interactive segmentation results with the different tools. Data on the left hand side shows a section from a abdominal MRI scan [31], the right hand side shows a section from a cervical CT scan [37].

zukunft.niedersachsen of the Volkswagen Foundation for the 'CAIMed – Lower Saxony Center for Artificial Intelligence and Causal Methods in Medicine' project (grant no. ZN4257). This work was also supported by the Google Research Scholarship "Vision Foundation Models for Bioimage Segmentation". We gratefully acknowledge the computing time granted by the Resource Allocation Board and provided on the supercomputer Emmy at NHR@Göttingen as part of the NHR infrastructure, under the project nim00007. We would like to thank Sebastian von Haaren for help with improving manuscript figures and data visualizations.

References

[1] Anwai Archit et al. "Segment Anything for Microscopy". In: *Nature Methods* (2025).
[2] Samir M. Badawy et al. "Automatic semantic segmentation of breast tumors in ultrasound images based on combining fuzzy logic and deep learning—A feasibility study". In: *PLoS ONE* (2021).
[3] Federico Bolelli et al. "Tooth fairy: A cone-beam computed tomography segmentation challenge". In: *MICCAI*. 2023.
[4] Tom Brown et al. "Language models are few-shot learners". In: *NeurIPS* (2020).
[5] Nhat-Tan Bui et al. "Sam3d: Segment anything model in volumetric medical images". In: *ISBI* (2024).
[6] M. Jorge Cardoso et al. "MONAI: An open-source framework for deep learning in healthcare". In: *arXiv* (2022).
[7] Fernando Cervantes-Sanchez et al. "Automatic segmentation of coronary arteries in X-ray angiograms using multiscale analysis and artificial neural networks". In: *Applied Sciences* (2019).
[8] Cheng Chen et al. "Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation". In: *Medical Image Analysis* (2024).
[9] Junlong Cheng et al. "Sam-med2d". In: *arXiv* (2023).
[10] Noel Codella et al. "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)". In: *arXiv* (2019).
[11] Tugba Akinci D'Antonoli et al. "TotalSegmentator MRI: Sequence-Independent Segmentation of 59 Anatomical Structures in MR images". In: *arXiv* (2024).
[12] Haoyu Dong et al. "Segment anything model 2: an application to 2d and 3d medical images". In: *arXiv* (2024).
[13] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ICLR*. 2021.
[14] J Glaister, A Wong, and D A. Clausi. "Automatic segmentation of skin lesions from dermatological photographs using a joint probabilistic texture distinctiveness approach". In: *IEEE Transactions on Biomedical Engineering* (2014).
[15] Karol Gotkowski. *napari-sam*. Github. 2020.
[16] Jasper W van der Graaf et al. "Lumbar spine segmentation in MR images: a dataset and a public benchmark". In: *Scientific Data* (2024).
[17] Titus Griebel, Anwai Archit, and Constantin Pape. "Segment Anything for Histopathology". In: *MIDL*. 2025.
[18] Hanxue Gu et al. "How to build the best medical image segmentation algorithm using foundation models: a compre-

hensive empirical study with Segment Anything Model". In: *MELBA* (2025).
[19] Ali Hatamizadeh et al. "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images". In: *arXiv* (2022).
[20] Ali Hatamizadeh et al. "UNETR: Transformers for 3D Medical Image Segmentation". In: *WACV* (2022).
[21] Kaiming He et al. "MAE: Masked autoencoders are scalable vision learners". In: *CVPR* (2022).
[22] Sheng He et al. "Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets". In: *arXiv* (2023).
[23] Nicholas Heller et al. "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge". In: *Medical Image Analysis* (2021).
[24] Fabian Hörst et al. "Cellvit: Vision transformers for precise cell segmentation and classification". In: *Medical Image Analysis* (2024).
[25] Edward J Hu et al. "Lora: Low-rank adaptation of large language models". In: *ICLR* (2022).
[26] Yuhao Huang et al. "Segment anything model for medical images?" In: *Medical Image Analysis* (2024).
[27] Fabian Isensee et al. "nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation". In: *MICCAI*. 2024.
[28] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* (2021).
[29] Stefan Jaeger et al. "Automatic tuberculosis screening using chest radiographs". In: *IEEE Transactions in Medical Imaging* (2013).
[30] Wei Ji et al. "Segment anything is not always perfect: An investigation of sam on different real-world applications". In: *Machine Intelligence Research* (2024).
[31] Yuanfeng Ji et al. "Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation". In: *NeurIPS* (2022).
[32] Hongxu Jiang et al. "MicroSegNet: A deep learning approach for prostate segmentation on micro-ultrasound images". In: *Computerized Medical Imaging and Graphics* (2024).
[33] Alexander Kirillov et al. "Segment anything". In: *ICCV*. 2023.
[34] Taha Koleilat et al. "MedCLIP-SAMv2: Towards Universal Text-Driven Medical Image Segmentation". In: *arXiv* (2024).
[35] Oleksandr Kovalyk et al. "PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment". In: *Scientific Data* (2022).
[36] Markus Krönke et al. "Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry". In: *PLOS One* (2022).
[37] Bennett Landman et al. "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge". In: *MICCAI*. 2015.
[38] Sarah Leclerc et al. "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography". In: *IEEE Transactions in Medical Imaging* (2019).
[39] Ho Hin Lee et al. "Foundation Models for Biomedical Image Segmentation: A Survey". In: *arXiv* (2024).
[40] Rebecca Sawyer Lee et al. "A curated mammography data set for use in computer-aided detection and diagnosis research". In: *Scientific Data* (2017).
[41] Wenhui Lei et al. "Medlsam: Localize and segment anything model for 3d medical images". In: *Medical Image Analysis* (2025).
[42] Chengyin Li et al. "MulModSeg: Enhancing Unpaired Multi-Modal Medical Image Segmentation with Modality-Conditioned Text Embedding and Alternating Training". In: *WACV* (2025).
[43] Yuheng Li, Mingzhe Hu, and Xiaofeng Yang. "Polyp-sam: Transfer sam for polyp segmentation". In: *Medical Imaging 2024: Computer-Aided Diagnosis*. SPIE. 2024.
[44] Haofeng Liu et al. "Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning". In: *NeurIPS Workshops* (2024).
[45] Jie Liu et al. "CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection". In: *ICCV*. 2023.
[46] Yihao Liu et al. "Samm (segment any medical model): A 3d slicer integration to sam". In: *arXiv* (2023).
[47] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *ICLR* (2019).
[48] Yaosheng Lu et al. "The JNU-IFM dataset for segmenting pubic symphysis-fetal head". In: *Data in Brief* (2022).
[49] Jun Ma et al. "Segment anything in medical images". In: *Nature Communications* (2024).
[50] Jun Ma et al. "Segment Anything in Medical Images and Videos: Benchmark and Deployment". In: *arXiv* (2024).
[51] Jacob A Macdonald et al. "Duke Liver Dataset: A publicly available liver MRI dataset with liver segmentation masks and series labels". In: *Radiology: AI* (2023).
[52] Caroline Magg et al. "Training-free Prompt Placement by Propagation for SAM Predictions in Bone CT Scans". In: *MIDL*. 2024.
[53] Salman Maqbool et al. "m2caiseg: Semantic segmentation of laparoscopic images using convolutional neural networks". In: *arXiv* (2020).
[54] Daniel S Marcus et al. "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults". In: *Journal of Cognitive Neuroscience* (2007).
[55] Maciej A Mazurowski et al. "Segment anything model for medical image analysis: an experimental study". In: *Medical Image Analysis* (2023).
[56] Cancer Genome Atlas Research Network. "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas". In: *New England Journal of Medicine* (2015).
[57] Antonio Pepe, Gian Marco Melito, and Jan Egger. "Segmentation of the Aorta: Towards the Automatic Segmentation, Modeling, and Meshing of the Aortic Vessel Tree from Multicenter Acquisition". In: *MICCAI* (2023).
[58] Gašper Podobnik et al. "HaN-Seg: The head and neck organ-at-risk CT and MR segmentation dataset". In: *Medical Physics* (2023).
[59] Maria Popa, Gabriela Adriana Vișa, and Ciprian Radu Șofariu. "PediMS: A Pediatric Multiple Sclerosis Lesion Segmentation Dataset". In: *Scientific Data* (2025).
[60] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *arXiv* (2021).
[61] Nikhila Ravi et al. "Sam 2: Segment anything in images and videos". In: *ICLR* (2025).
[62] Meritxell Riera-Marin et al. "Calibration and Uncertainty for multiRater Volume Assessment in multiorgan Segmentation (CURVAS) challenge results". In: *arXiv* (2025).
[63] Julio César Mello Román et al. "Panoramic dental radiography image enhancement using multiscale mathematical morphology". In: *Sensors* (2021).
[64] Luisa F Sánchez-Peralta et al. "Piccolo white-light and narrow-band imaging colonoscopic dataset: A performance comparative of models and datasets". In: *Applied Sciences* (2020).
[65] Ahmed Shahin et al. *OSIC Pulmonary Fibrosis Progression*. Kaggle. 2020.
[66] Carolin Teuber, Anwai Archit, and Constantin Pape. "Parameter Efficient Fine-Tuning of Segment Anything Model for Biomedical Imaging". In: *MIDL*. 2025.
[67] Ashish Vaswani et al. "Attention is all you need". In: *NeurIPS* (2017).
[68] Haoyu Wang et al. "SAM-Med3D". In: *ECCV Workshops* (2024).
[69] Jakob Wasserthal et al. "TotalSegmentator: robust segmentation of 104 anatomic structures in CT images". In: *Radiology: AI* (2023).

[70] Junde Wu et al. “Medical sam adapter: Adapting segment anything model for medical image segmentation”. In: *Medical Image Analysis* (2025).
[71] Xinyu Xiong et al. “SAM2-UNet: Segment Anything 2 Makes Strong Encoder for Natural and Medical Image Segmentation”. In: *arXiv* (2024).
[72] Yanwu Yang et al. “MedSAMix: A Training-Free Model Merging Approach for Medical Image Segmentation”. In: *arXiv* (2025).
[73] Jin Ye et al. “Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks”. In: *arXiv* (2023).
[74] Xin Ye et al. “OIMHS: An Optical Coherence Tomography Image Dataset Based on Macular Hole Manual Segmentation”. In: *Scientific Data* (2023).
[75] Zafer Yildiz et al. “SegmentWithSAM: 3D Slicer Extension for Segment Anything Model (SAM)”. In: *MIDL*. 2024.
[76] Chongkai Yu et al. “SAM-REF: Introducing Image-Prompt Synergy during Interaction for Detail Enhancement in the Segment Anything Model”. In: *arXiv* (2024).
[77] Wenxi Yue et al. “Surgicalsam: Efficient class promptable surgical instrument segmentation”. In: *AAAI*. 2024.
[78] Anna Zawacki et al. *SIIM-ACR Pneumothorax Segmentation*. Kaggle. 2019.
[79] Kaidong Zhang and Dong Liu. “Customized segment anything model for medical image segmentation”. In: *arXiv* (2023).
[80] Yichi Zhang, Zhenrong Shen, and Rushi Jiao. “Segment anything model for medical image segmentation: Current applications and future directions”. In: *Computers in Biology and Medicine* (2024).
[81] Theodore Zhao et al. “A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities”. In: *Nature Methods* (2024).
[82] Ziheng Zhao et al. “Large-Vocabulary Segmentation for Medical Images with Text Prompts”. In: *arXiv* (2023).
[83] Jiayuan Zhu, Yunli Qi, and Junde Wu. “Medical sam 2: Segment medical images as video via segment anything model 2”. In: *arXiv* (2024).