INTRABENCH:
INTERACTIVE RADIOLOGICAL BENCHMARK

Constantin Ulrich * [1,4,5], **Tassilo Wald*** [1,2,3], **Emily Tempus** * [1],
Maximilian Rokuss[1,3], **Paul F. Jaeger**[2,6,7], **Klaus Maier-Hein**[1,2,3,4,5,7]
[1] Division of Medical Image Computing, German Cancer Research Center (DKFZ)
[2] Helmholtz Imaging, DKFZ, Heidelberg, Germany
[3] Faculty of Mathematics and Computer Science, University of Heidelberg, Germany,
[4] Medical Faculty Heidelberg, University of Heidelberg, Germany
[5] National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and university medical center Heidelberg
[6] Interactive Machine Learning Group, DKFZ Heidelberg, Germany
[7] Pattern Analysis and Learning Group, Department of Radiation Oncology
constantin.ulrich@dkfz-heidelberg.de

ABSTRACT

Current interactive segmentation approaches, inspired by the success of META's Segment Anything model, have achieved notable advancements, however they come with substantial limitations that hinder their practical application in real clinical scenarios. These include unrealistic human interaction requirements, such as slice-by-slice operations for 2D models on 3D data, a lack of iterative refinement, and insufficient evaluation experiments. These shortcomings prevent accurate assessment of model performance and lead to inconsistent outcomes across studies.

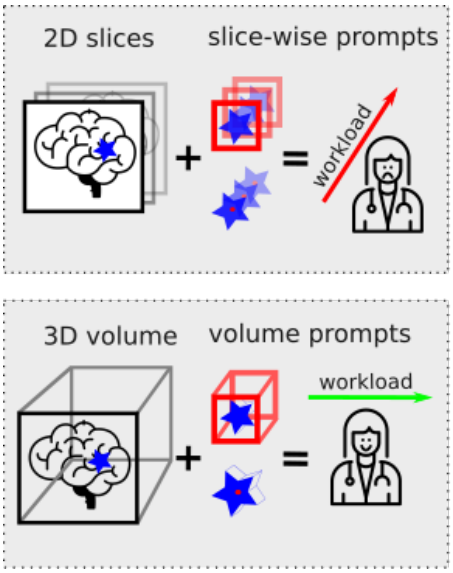
`IntRaBench` overcomes these challenges by offering a comprehensive and reproducible framework for evaluating interactive segmentation methods in realistic, clinically relevant scenarios. It includes diverse datasets, target structures, and segmentation models, and provides a flexible codebase that allows seamless integration of new models and prompting strategies. Additionally, we introduce advanced techniques to minimize clinician interaction, ensuring fair comparisons between 2D and 3D models. By open-sourcing IntRaBench, we invite the research community to integrate their models and prompting techniques, ensuring continuous and transparent evaluation of interactive segmentation models in 3D medical imaging.

1 INTRODUCTION

Accurate segmentation of anatomical structures or pathological areas is crucial in fields like radiology, oncology, and surgery to isolate affected regions, monitor disease progression, treatment planning and guide therapeutic procedures. Traditional supervised medical segmentation models have demonstrated strong performance across a range of anatomies and pathologies (Isensee et al., 2020; 2023; Huang et al., 2023; Ulrich et al., 2023). However, their effectiveness remains heavily constrained by the amount and diversity of available training data, with the quality of human label annotations serving as a critical limiting factor. Consequently, fully autonomous AI solutions have not yet reached performance needed for widespread autonomous clinical applications.

On the other hand, numerous semi-automatic segmentation techniques, not reliant on AI, are already in clinical practice to expedite manual annotation processes Hemalatha et al. (2018). These current ad hoc methods do not tap into the potential of AI-based automation to drastically reduce annotation time. A method that allows clinicians to segment any target with just a single click within the image could greatly enhance the efficiency of clinical workflows.
The release of META's Segment Anything (SAM) model represents a big leap towards making this

*Equal contribution. Authors are permitted to list their name first in their CVs.



	Dim	Point	Box	Mask
SAM		✓	□	✓
SAM2		✓	□	✓
MedSAM		✗	□	✗
SAM-Med 2D		✓	□	✓
SAM-Med 3D		✓	✗	✓
SAM-Med 3D Turbo		✓	✗	✓
SegVol		✓		✗

Figure 1: a) Current approaches require clinicians to interact with radiological images slice by slice, leading to increased workload. b) Some models operate natively in 3D and enable full 3D interaction. Only models that accept mask prompts allow iterative refinement of initial predictions with human guidance.

potential a reality (Kirillov et al., 2023). "SAM" is designed to segment any target through different user interaction methods, including point-based and bounding box prompts. This allows users to easily specify the area of interest by clicking on it or drawing a bounding box around it, making the segmentation process both flexible and intuitive. A particularly powerful feature is the ability for users to iteratively refine initial predictions by adding more positive or negative prompts.

This advanced functionality, in contrast to traditional supervised segmentation methods, has attracted a lot of attention in the medical domain, and led to many studies evaluating and adapting SAM for 3D medical image segmentation (Roy et al., 2023; Deng et al., 2023; Hu et al., 2023; Zhou et al., 2023; Mohapatra et al., 2023; Cheng et al., 2023; Ma et al., 2024; Gong et al., 2023). Moreover, several researchers have been inspired by SAM's capabilities to develop their own methods, often specifically designed for the 3D nature of radiological data (Du et al., 2024; He et al., 2024; Li et al., 2024; Wang et al., 2024).

Although these domain-specific adaptations on medical data have shown promising progress, many published methods are plagued by pitfalls which obfuscate the efficacy of the models and prevent clinicians and researchers from determining the best methods for their use-cases:

Applying interactive 2D models to 3D data on a slice-by-slice basis (P1): Assuming clinicians will interact with each slice individually is unrealistic and undermines the efficiency improvements these methods aim for. Moreover, a slice-by-slice approach introduces an unfair bias when comparing 2D and 3D models, as 3D models typically require only a few interaction per image, leading to significantly fewer interactions and less supervision Cheng et al. (2023); Ma et al. (2024); Zhang & Liu (2023); Wu et al. (2024); Wong et al. (2024).

Neglecting refinement (P2): Many studies assess interactive segmentation methods based on a single interaction step, overlooking the inherent ambiguities in radiological images (Ma et al., 2024; Du et al., 2024; Gong et al., 2023; Bui et al., 2024). Often, a second interaction may be necessary to specify which specific substructure the clinician wants to segment. This could be, e.g. a vessel within the liver, or the necrosis within a tumor, as exemplified in the well-known BraTs segmentation challenge (de Verdier et al., 2024). Furthermore, clinicians often want to adapt the segmentations to their clinic's local protocol or refine them, particularly for targets with high inter-rater variability, like pathological structures (Fu et al., 2014; Benchoufi et al., 2020; Hesamian et al., 2019). Overall, there is a notable lack of research exploring realistic, iterative interaction methods for 2D models applied to 3D volumes.

Obfuscated and insufficient evaluation (P3): With promptable models only recently garnering great attention, there is a lack of a standardized approach to evaluation, which has led to disparate and incomparable methods, which are at times even obfuscated or insufficient. We observed the following shortcomings: (i) Not specifying whether predictions were interactively refined or based on a single prompt with multiple points (Cheng et al., 2023; Wang et al., 2024). (ii) Being intransparent on the number of initial prompts given (Du et al., 2024). (iii) Using the best mask rather than the final mask after interactive refinement (Wang et al., 2024). (iv) Evaluating predictions slice-by-slice or on sub-patches of a 3D volume instead of evaluating on the full image (Roy et al., 2023; Ma et al., 2024; Cheng et al., 2023; He et al., 2024; Li et al., 2024). (v) Excluding targets considered 'too small', hence neglecting valid targets such as small lesions that are neither tested nor trained on Ma et al. (2024); Cheng et al. (2023); Wang et al. (2024). (vi) Comparing against non-promptable models and SAM, rather than any other promptable model trained on medical data (Cheng et al., 2023; Ma et al., 2024; Gong et al., 2023; He et al., 2024). (vii) Lastly, overemphasizing segmenting healthy structures, such as organs, where existing supervised public models already perform well (Wasserthal et al., 2023; Ulrich et al., 2023), instead of focusing on pathologies, where interactive refinement could provide the greatest benefits (Wang et al., 2024; Zhang & Liu, 2023).

To address these pitfalls, a benchmark is needed, aligning with the recent review paper from Marinov et al. (2024). To this end, we introduce `IntRaBench`, a reproducible and extendable Interactive Radiological Benchmark. Through it, we highlight the most performant 2D and 3D interactive segmentation and the best prompting methods in the radiological domain. In this paper, we present experiments carefully designed to replicate a clinical workflow as closely as possible, with the following key contributions:

1. `IntRaBench`, for the first time, enables a fair comparison of the most influential 2D and 3D interactive segmentation methods. By measuring the number of simulated interactions, a proxy for the 'Human Effort', we test different prompting strategies that do not require a slice-wise interaction (P1).
2. We propose effective interaction strategies for refinement of predictions in a 3D volume, without requiring clinicians to interact with each individual slice (P2).
3. We provide a standardized evaluation protocol to generate prompts, select model outputs and compute the segmentation metrics on the entire image across ten datasets, covering various modalities and target structures, including small lesions (P3). Our benchmarking efforts include a performance comparison against leading interactive segmentation methods in the medical domain.
4. The extendable `IntRaBench` framework allows developers to a) easily evaluate a new method in a fair manner against established methods and b) easily develop and investigate new prompting strategies.

Through open-sourcing `IntRaBench`, we invite researchers to integrate their methods into our framework, promoting continuous and equitable assessment that allows us to track the overall progress in the field of interactive 3D medical image segmentation reproducibly and transparently.

2 INTRABENCH

The Interactive Radiology Benchmark is designed to easily enable a fair and reproducible evaluation of 2D and 3D interactive segmentation methods for 3D radiological image segmentation for the very first time. While prompting 3D models is generally straightforward, we introduce specific prompting and refinement strategies for 2D models to streamline human interaction and reduce the simulated 'Human Effort'. The proposed benchmark includes seven established models and ten datasets covering different target structures and image modalities. All datasets are publicly available and we support an automatic download and preprocessing for improved usability and reproducibility.

Moreover, the benchmark is built with flexibility in mind, enabling seamless integration of additional methods, as visualized in Fig. 2. Researchers are invited to contribute new approaches, particularly new models, new prompting schemes, and new interesting datasets to the collection. Overall, the design of our benchmark allows for easy testing and validation of novel segmentation methods, making the benchmark a catalyst for advancing methodology for interactive 3D medical image segmentation. In the following, we present the different components of `IntRaBench`.

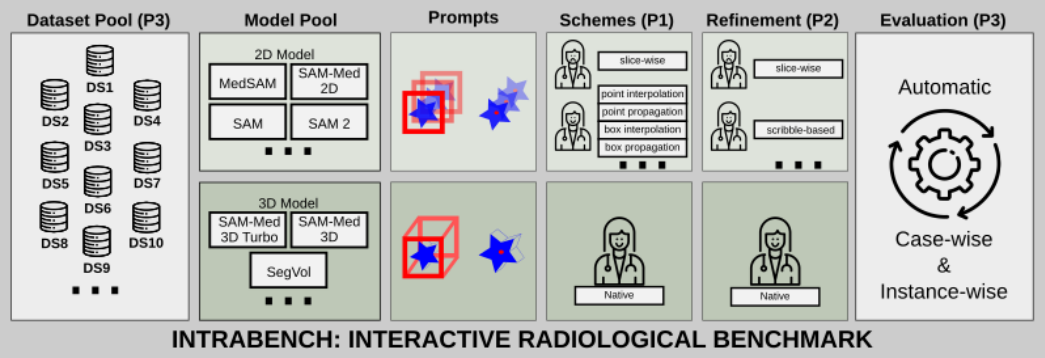


Figure 2: `IntRaBench` overview. Although our evaluation is performed on entire 3D volumes, the benchmark accommodates both 3D and 2D interactive segmentation methods. While 3D model prompting is relatively straightforward, we introduce prompting and refinement strategies for 2D models that minimize the effort required from human interaction. The benchmark is designed to be extensible, and researchers are encouraged to propose and integrate additional methods seamlessly using our codebase particularly for areas marked by three dots.

2.1 Initial Prompting

Prompts are a key component of any interactive segmentation method and can highly influence the overall performance of the underlying method. `IntRaBench`distinguishes between two visual prompting types. **Point prompts** correspond to a click of a user in the image, and **box prompts** refer to a box around the target structure. While there is no difference in providing a point prompt for 2D and 3D methods, a 3D box requires an additional dimension compared to a 2D box. Notably, some methods also enable a distinction between foreground and background point prompts. While 3D models allow segmenting a 3D volume natively, 2D-based models require an interaction for each slice, resulting in excessive effort, which is prohibitive for clinicians as it would take too much time in daily clinical practice. Hence, any meaningful performance comparison must account for this difference in prompting effort.

To increase the feasibility of 2D models for 3D applications, it is essential to reduce this effort. We propose two straightforward methods, for both point and box prompts, to explore their performance and provide a proxy for measuring the effort of human interaction.

Point interpolation: Let $I \subset N$ be a set of axial indices of all foreground slices. We simulate a user by selecting n foreground points, specifically the center of the largest connected component of slice $i_1, ...i_n \in I$ where the i_j are equally spaced within I and $i_1 = \min(I)$ and $i_n = \max(I)$. Then, we interpolate linearly between each point and the next one and use the intersections of the resulting lines with the axial slices as positive point prompts, as visualized in Fig. 3 c).

Point propagation: We simulate a user providing $\min(I)$, $\max(I)$, and a 2D point prompt within the median slice corresponding to the median axial index i_m. Given this point, the model generates a segmentation S_m for the median slice. We then calculate a 'central point,' specifically the center of mass of the largest connected component of S_m, to use as a point prompt for the slice indexed by i_{m-1}. Again, we generate a segmentation S_{m-1} of this slice, and create a new central point until we segment the slice with the axial index $\min(I)$. The propagation is then repeated upwards, starting from i_{m+1} and continuing until we segment the slice with the axial index $\max(I)$. This process is visualized in Fig. 3 e).

Box interpolation: We simulate a user providing n 2D bounding boxes, one in each of $i_1, ..., i_n \in I$, with i_j defined as in the point interpolation paragraph. Since the boxes are uniquely defined by their minimum and maximum vertices, we can interpolate between the minimum vertices as in point propagation to get a minimum vertex in each axial slice, and similarly get a maximum vertex in each axial slice, providing a box prompt in each slice. This box interpolation is exemplified in Fig. 3 d).

Box propagation: We simulate a user providing $\min(I)$, $\max(I)$, and a 2D box prompt within the slice corresponding to the axial index i_m, m as in point propagation. The model then generates

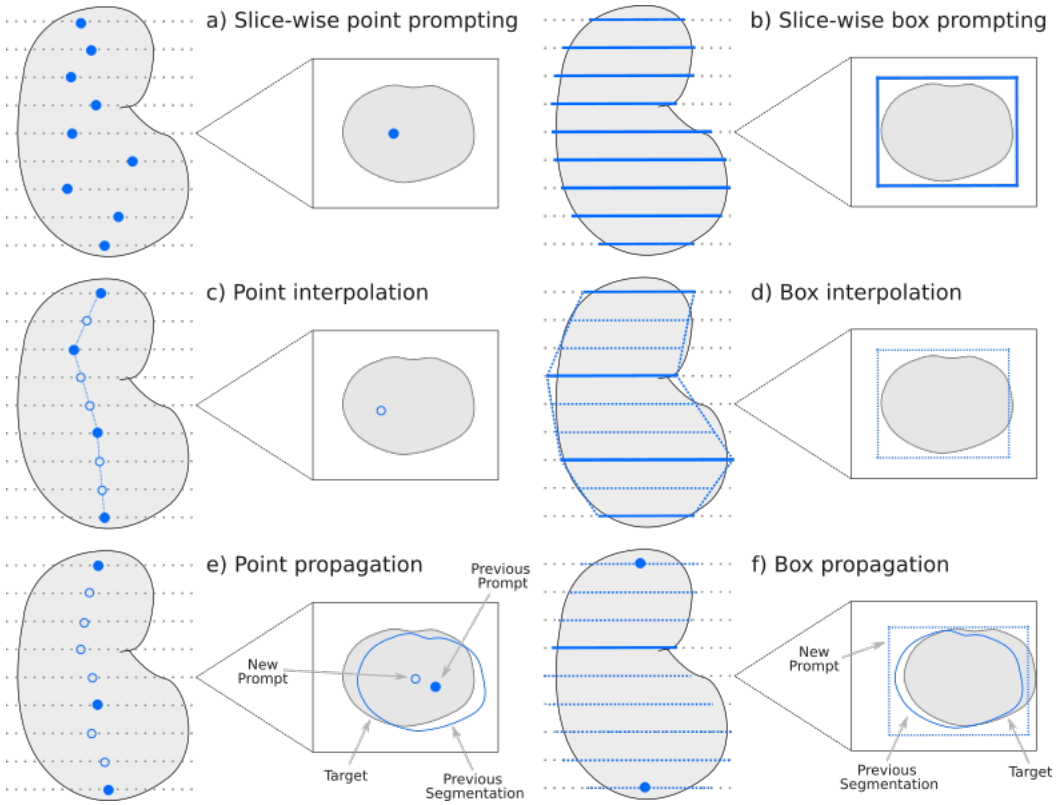


Figure 3: Different promoting schemes for 2D models based on point prompts (on the left) and box prompts (on the right). While a) and b) expect unrealistic human slice-by-slice interaction, c) and d) illustrate the proposed prompt interpolation schemes, where a human needs to provide prompts for at least 3 slices (4 slices in this case). Prompts for the remaining slices are generated by interpolating between the initial prompts. e) and f) present the proposed prompt propagation methods, where the prompt for each subsequent slice is automatically generated based on the model's prediction from the previous slice. Only the initial slice and upper and lower boundaries require manual prompts.

a segmentation S_m for the median slice. A new bounding box is created based on S_m and used as a prompt for the slice indexed by $i_m - 1$. The propagation is continued down to $\min(I)$ and then repeated upwards until $\max(I)$ as in point propagation, but using box prompts instead of point prompts. See Fig. 3 f) for a visualization.

While one cares about realistic prompting behavior, `IntRaBench` also supports the previously mentioned slice-by-slice prompting styles for completeness.

2.2 Refinement Prompting

Refinement of previous segmentations is an important aspect of interactive segmentation models, as it allows iteratively improving the segmentation until the desired structure is segmented to a user's demands. Some interactive segmentation models allow for the refinement of initial segmentations by providing the model with the previous prediction along with a new prompt to correct errors, either through foreground clicks on false negative pixels or background clicks on false positive pixels. While this process is straightforward for 3D models, 2D models naively only allow for refinement in a slice-by-slice fashion, which again places an unrealistic burden on clinicians. Therefore, we present refinement strategies that require a reasonable level of "Human Effort".

Scribble refinement: To represent a user-centric refinement strategy we introduce an algorithm simulating user-created scribble prompts: At each refinement step, our proposed algorithm generates either positive or negative additional prompts. The decision to generate positive prompts follows a

Bernoulli trial with success probability $p = n_{fn}/(n_{fn}+n_{fp})$, where n_{fn}, n_{fp} represent the number of false negatives and false positive voxels, respectively.
If positive prompts are selected, we perform a connected component analysis on the false negative voxels. Given L, the largest connected component, we generate a scribble from the bottom to the top of L by taking the centroid of L in each slice to simulate drawing a vertical scribble through the 'middle' of L. This simulates a clinician annotating regions that were erroneously not segmented.
For 2D models, we then individually feed all slices $i \in I$ where the voxel along the scribble was not predicted, along with the new positive prompt derived from the scribble and the previous prediction $s \subset S$, back into the model. For 3D models, we feed the whole 3D patch, together with the previous prediction S, and all new positive points derived from the scribble into the network in one step.
If negative prompts are selected, we identify a non-axial slice S_{fp} of S that contains the most false positives. Then we generate a contour curve around the ground truth target object at a distance of 2 pixels. We then select a subpart C with a length of 60% of the full curve and sample all pixels $c \in C$ that are false positives to obtain a set of points D, simulating a user drawing a few scribbles in areas where the model over-segmented the target. For 2D models, we then generate new slice predictions for each slice containing a point in D by providing the model with the previous prediction as well as new negative prompts: all $d \in D$ which belong to that slice. For 3D models, we again feed the whole 3D patch, together with the previous prediction S and a negative prompt sampled from D.

2.3 Human Effort Proxy

A model's performance is highly dependent on the effort a human puts into initial prompting and refinement of the predicted masks. Generally, the effort required for 3D methods is less than that for 2D methods, although the strategies mentioned above significantly reduce the effort of 2D methods substantially. We aimed to establish a general measure of the effort a method would require from a human user. A more formalized mathematical approach involves assigning degrees of freedom (DoF) to each interaction. For instance, a point corresponds to 3 DoF, a 2D box has 5 DoF (requiring selection of the z-axis and two 2D points), and a 3D box consists of 6 DoF. However, point interpolation has 9 DoF, whereas point propagation only has 5 DoF, since it requires just the axial coordinate rather than both minimum and maximum points with 3 DoF each. From the user's perspective, however, identifying the z-coordinate demands the same level of effort as selecting a 3D coordinate by clicking at the target structure's endpoint along the z-axis. Similarly, an arbitrary scribble has significantly more DoF than a straight or parabolic line, yet the difference in effort for the user is minimal. Therefore, we define user effort in terms of the number of interactions required for a specific task. While not an exact measure, this method offers the most practical estimation of the actual effort involved from the user's perspective.

2.4 Interactive methods

In our comprehensive benchmark, we include various interactive segmentation methods. Fig. 1 illustrates the types of prompts each method supports. Iterative refinement is only possible for methods that allow a (previously predicted) mask as a prompt.

SAM is the most prominent model from the natural image domain, that inspired many researchers to evaluate and adapt it to the domain of radiological medical images. It was trained on iteratively generated and curated 1B masks and 11M images, but not explicitly on radiological images. META's Segment Anything Model was the first to popularize interactive segmentation models (Kirillov et al., 2023).

SAM2 is an extension of SAM that was trained on even more images and introduced support for video data (Ravi et al., 2024).

MedSAM is an adaptation of SAM that fine-tuned SAM's weights on 1,570,263 image-mask pairs from the medical domain. It supports only a single forward pass without refinement and is limited to box prompts (Ma et al., 2024).

SAM-Med 2D is another adaptation of SAM, fine-tuned on 4.6 million images with 19.7 million masks from the medical domain. Unlike MedSAM, it supports points, boxes, and mask prompts, allowing for refinement (Cheng et al., 2023).

Table 1: Overview over all Datasets. None of these datasets were part of the original training data for the methods, except for SegVol, which utilized D2 HanSeg.

Dataset	Modality	Targets	Images
D1 MS Lesion (Muslim et al., 2022)	MRI (T2 Flair)	MS Lesions	60
D2 HanSeg (Podobnik et al., 2023)	MR (T1)	30 Organs at risk	42
D3 HNTSRMFG (Wahid et al., 2024)	MRI (T2)	Oropharyngeal cancer & metastatic lymph nodes	135
D4 RiderLung (Zhao et al., 2015)	CT	Lung lesions	58
D5 LNQ (Dorent et al., 2024)	CT	Mediastinal lymph nodes	513
D6 LiverMets (Simpson et al., 2023)	CT	Liver metastases	171
D7 Adrenal ACC (Moawad et al., 2023)	CT	Adrenal tumors	53
D8 HCC Tace (Moawad et al., 2021)	CT	Liver, Liver tumors	65
D9 Pengwin (Liu et al., 2023)	CT	Bone fragments	100
D10 Segrap (Luo et al., 2023)	CT	45 Organs at risk	30

SAM-Med 3D incorporates a transformer-based 3D image encoder, 3D prompt encoder, and 3D mask decoder. It was trained from scratch using 22,000 3D images and 143,000 corresponding 3D masks and supports point and mask prompts and also allows for refinement (Wang et al., 2024).

SAM-Med 3D Turbo is an updated version of SAM-Med 3D trained on a larger dataset collection of 44 datasets for improved performance. It supports the same prompt styles as SAM-Med 3D (Wang et al., 2024).

SegVol is an interactive 3D segmentation model based on a 3D adaptation of a ViT (Dosovitskiy, 2020) that was trained on 96K unlabelled CT images and fine-tuned with 6K labeled CT images. It supports points and bounding boxes as spatial prompts but does not allow iterative refinement (Du et al., 2024).

Aside from these models there exist other notable interactive models, such as Vista3D (He et al., 2024), 3D Sam Adapter (Gong et al., 2023) and Prism (Li et al., 2024). However, while being promptable, they are closed-set, i.e. not trained to segment any arbitrary prompted structure. Subsequently, they were not considered for this benchmark.

2.5 DATASETS

Dataset selection was a non-trivial problem for this benchmark: While models that were originally introduced in the natural image domain rarely see any radiological 3D data, the medical counterparts were often trained on all publicly available datasets that the authors could obtain. For example, MedSAM was trained using more than 60 publicly available datasets (Ma et al., 2024). Although these methods conducted their final validation on excluded datasets or at least on separate test subsets of images, the test datasets vary between models. As a result, identifying annotated datasets with interesting target structures that were not part of any of the included methods' training datasets has proven challenging.

Nevertheless, we assembled a diverse collection of ten lesser-known or recently released public datasets featuring various pathologies and organs, including CT and MRI image modalities. Specific details of these are provided in Table 1. To enhance reproducibility and eliminate barriers of entry for non-domain experts, we automated the dataset download and preprocessing, minimizing any required domain knowledge to use the benchmark. However, due to the sparsity of labeled datasets, we urge developers to exclude these datasets from their train dataset selection, as inclusion would compromise the integrity of a clean evaluation through `IntRaBench`.

2.6 EVALUATION

All interactive segmentation methods identify their target structure based on a spatial prompt, inherently resulting in instance segmentation. As a result, we evaluate on an instance-by-instance basis. Unlike in object detection, each prompt already provides information on the localization of the target structure, making detection metrics like F1-Score irrelevant. Subsequently, we rely solely on the Dice Similarity Coefficient (DSC) score as a metric. The instance-wise DSC metric is then averaged per case (i.e. per image volume), and further aggregated across all cases in the dataset, as

recommended by Maier-Hein et al. (2024). For better presentation, we averaged the DSC across all classes of a dataset and also specified how many human interactions are simulated.

3 EXPERIMENTS

We evaluate all seven models across various initial prompting scenarios under realistic and unrealistic effort settings. Following this, we conduct interactive experiments to simulate human refinement of model predictions. Due to the vast amount of data, we only provide a condensed version of the results for easier insights. Detailed results and the number of human interactions are provided in the Appendix B.

3.1 INITIAL PREDICTION

Unrealistic effort: As an upper baseline, we begin with an idealized and unrealistic scenario where each slice is prompted individually for all 2D models. In this setting, we evaluate different numbers of point prompts per slice (PPS), as well as alternating positive and negative prompts ($\pm$ PPS), and slice-wise box prompts with varying numbers of boxes per slice (BPS). Figure 4 shows that models employing box prompts achieved significantly higher average Dice scores, with SAM2 demonstrating the strongest performance across all models. Conversely, point-based prompts performed poorly, particularly for small target regions, such as small MS lesions in dataset D1 (see Appendix B.1). SAM Med2D outperforms non-medical models for point prompts. Although including positive and negative prompts and increasing the number of point prompts led to improvements, these were minor compared to the marked superiority of box-based prompts. These results highlight the limitations of point prompts, especially in cases involving small or complex anatomical structures, and emphasize the robustness of box prompts in achieving higher segmentation accuracy.

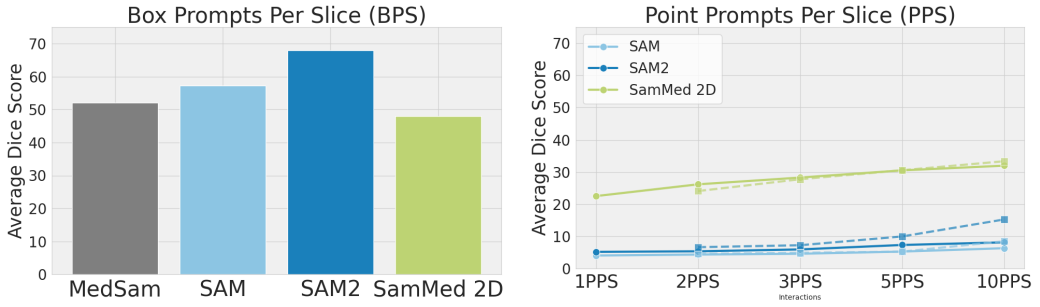


Figure 4: **Unrealistic prompting of 2D Boxes each slice performs best.** When comparing model's prompted with one Box Prompts Per Slice (BPS) (left) with various Point Prompts per Slice (PPS) (right) boxes perform better. While alternating positive and negative points (dashed lines) is slightly superior to only positive points the gap between points and boxes remains large.

Realistic Effort: To simulate a human-in-the-loop scenario, we evaluate various prompting strategies that avoid slice-by-slice interaction. As described in Section 2, for 2D models, we test point and box interpolation, as well as propagation, using different numbers of initial prompts. For 3D models, we explore varying numbers of Point prompts Per Volume (PPV) and 3D box prompts. Fig. 5 presents the following key findings:

1. For all models, box interpolation with 3 or 5 initial 2D boxes is sufficient to achieve results similar to slice-wise box prompting (BPS).
2. For SAMMed 2D, using 3 points with simple point interpolation achieves results comparable to prompting every slice.
3. SAM 2 outperforms specialized medical models across all prompting schemes using box interpolation.
4. Among 3D models, only SegVol is competitive to 2D models that use box prompts.

5. Both box and point propagation perform worse than their interpolation counterpart, though this may improve as models evolve.

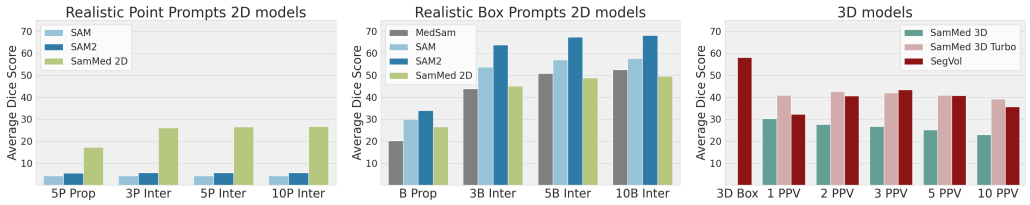


Figure 5: **Simple Interpolation Strategies Match Unrealistic Slice-Wise Prompting.** Sampling prompts from the interpolated connection between three initial prompts yields similar performance for SAM2 as slice-wise prompting (left). This is also observed for box interpolation across all models (middle). 3D models perform worse than 2D methods when only a few points are provided, while SegVol demonstrates that using a 3D box is superior to points (right).

3.2 Interactive Refinement

Finally, we evaluate the performance of the models during iterative refinement. For 2D models, this involves prompting on a slice-by-slice basis. As illustrated in Fig. 6 (left), adding refinement prompts to each slice results in a substantial performance boost. Although the proposed scribble-based refinement consistently improves outcomes, it does not achieve the same level of improvement as adding a prompt to every slice, which is expected since not all slices receive new prompts during the scribble refinements. We observed that for 2D models, it is crucial to provide the initial prompts again for each of the refinement steps. 2D models tend to over-segment the target, filling the entire slice foreground. The absence of the initial prompt leads to a complete loss of target location information, as the initial predicted mask is highly inaccurate. Our refinement likely generates negative additional prompts due to the large number of false positive pixels. In Table 5, we present refinement results from initial predictions produced by Box Interpolation. In this case, we did not include the previous point in the iterative prompts, which resulted in a performance decline during refinement.

For 3D models, iterative refinement also led to consistent performance improvements. Both randomly sampled prompts and those derived from refinement scribbles improved performance with each refinement iteration. Although SegVol initially performs best during initial prediction, it lacks support for further refinement. In contrast, SamMed 3D Turbo - worse in the initial prediction - surpasses SegVol, which was prompted using points, after several refinement steps.

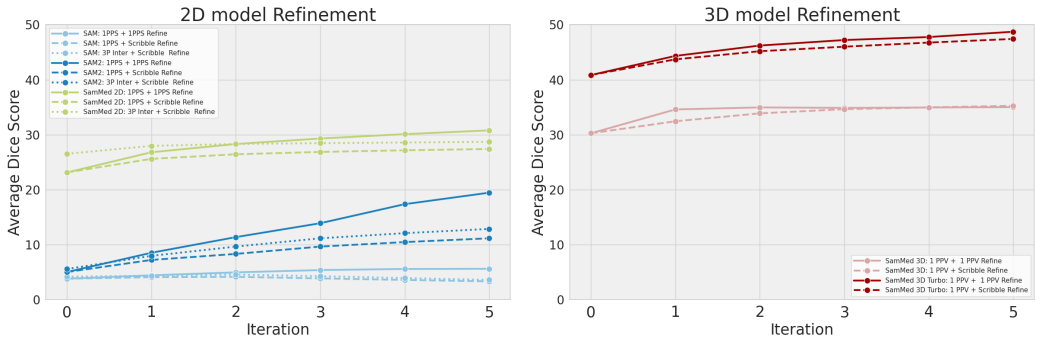


Figure 6: **All models demonstrate significant improvements from iterative refinement**. Results for 2D models are shown on the left, and for 3D models on the right. Dashed lines represent the use of the proposed scribble refinement. While the unrealistic scenario of one refinement point per slice yields better performance, the proposed scribble refinement consistently enhances results across iterations for 2D models.

3.3 Discussion and Conclusion

In this paper we introduced IntRaBench and with it, compared the performance of 2D and 3D interactive segmentation models in 3D medical imaging. We provide a holistic and transparent overview of the current state-of-the-art and highlight key findings that offer practical insights:

1. **Bounding Boxes Outperform Points:** Bounding boxes consistently outperform point-based inputs by providing better spatial context, which leads to improved segmentation accuracy, especially for complex structures in radiological images. Point-based prompts lack this context, resulting in poorer performance.
2. **Iterative Refinement is Essential:** The ability to iteratively refine segmentations significantly enhances model performance, particularly in challenging cases. Models that allow multiple rounds of corrections show better accuracy, making this feature crucial for clinical applications. For example SegVol reached highest performance in a static setting, however SamMed 3D Turbo is able to exceed SegVol given a few interactions, highlighting the importance of refinement.
3. **Realistic 2D prompting can match unrealistic prompting:** Our introduced realistic prompting styles are able to reach and match unrealistic prompting 2D prompting methods, see Fig. 5. This unlocks 2D methods for actual clinical workflows without any performance penalties.
4. **Points fail for difficult and small structures:** Contrary to claims in previous literature, point-based methods fail, likely due to previous work training and evaluating their methods on simpler target structures.

Implications IntRaBench suggests that bounding boxes and iterative refinement should be prioritized in the design of segmentation models for medical imaging, particularly when addressing complex radiological images. Furthermore, it underscores the importance of including diverse, difficult tasks in training data to improve model generalization for clinical use. It is also crucial to test 2D models in scenarios that simulate real human interaction, ensuring that segmenting a volumetric image does not require unreasonable effort by prompting the model slice-by-slice.

A key limitation of this work is that it only simulates "Human Effort". While this approach provides valuable insights into model performance by providing a proxy for the simulated "Human Effort", it falls short of capturing the full complexity and practical challenges of real clinical applications. As future work, a comprehensive study involving clinicians is essential to assess different prompting strategies in real-world environments. Such a study should not only evaluate segmentation performance but also measure the time required for annotation, offering critical insights into the practical feasibility and efficiency of these models in clinical practice.

To conclude, our proposed IntRaBench presents a powerful tool for the future of interactive segmentation research in medical imaging, serving as a catalyst for innovative solutions by enabling a fair and reproducible comparison between leading methods. One of the standout potentials is its ability to streamline the evaluation of both 2D and 3D segmentation models, allowing for more realistic and clinically relevant testing conditions. By focusing on human interaction and the efficiency of iterative refinement, IntRaBench opens new avenues for research, including understanding the impact of different interaction strategies and how they reduce clinician effort. Not only does this benchmark address the existing gaps in evaluation standardization, but it also offers a unique opportunity to refine segmentation performance on pathologies often overlooked, such as small lesions. The open-source nature of the benchmark further encourages continuous contributions, allowing researchers to test new methods and prompting strategies seamlessly within this framework. Future work using IntRaBench can reveal novel insights into the balance between performance and clinician involvement. This potential to improve real-world clinical applications, especially by reducing the labor intensity of medical professionals, marks IntRaBench as a crucial tool in catalyzing meaningful research progress.

REFERENCES

M. Benchoufi, E. Matzner-Lober, N. Molinari, A.-S. Jannot, and P. Soyer. Interobserver agreement issues in radiology. *Diagnostic and Interventional Imaging*, 101(10):639–641, October 2020. ISSN 2211-5684. doi: 10.1016/j.diii.2020.09.001. URL http://dx.doi.org/10.1016/j.diii.2020.09.001.

Nhat-Tan Bui, Dinh-Hieu Hoang, Minh-Triet Tran, Gianfranco Doretto, Donald Adjeroh, Brijesh Patel, Arabinda Choudhary, and Ngan Le. Sam3d: Segment anything model in volumetric medical images, 2024. URL https://arxiv.org/abs/2309.03493.

Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, Hui Sun, Junjun He, Shaoting Zhang, Min Zhu, and Yu Qiao. Sam-med2d, 2023. URL https://arxiv.org/abs/2308.16184.

Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, Ken Chang, and et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri, 2024.

Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W. Remedios, Shunxing Bao, Bennett A. Landman, Lee E. Wheless, Lori A. Coburn, Keith T. Wilson, Yaohong Wang, Shilin Zhao, Agnes B. Fogo, Haichun Yang, Yucheng Tang, and Yuankai Huo. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging, 2023. URL https://arxiv.org/abs/2304.04155.

Reuben Dorent, Roya Khajavi, Tagwa Idris, Erik Ziegler, Bhanusupriya Somarouthu, Heather Jacene, Ann LaCasce, Jonathan Deissler, Jan Ehrhardt, Sofija Engelson, Stefan M. Fischer, Yun Gu, Heinz Handels, Satoshi Kasai, Satoshi Kondo, Klaus Maier-Hein, Julia A. Schnabel, Guotai Wang, Litingyu Wang, Tassilo Wald, Guang-Zhong Yang, Hanxiao Zhang, Minghui Zhang, Steve Pieper, Gordon Harris, Ron Kikinis, and Tina Kapur. Lnq 2023 challenge: Benchmark of weakly-supervised techniques for mediastinal lymph node quantification, 2024. URL https://arxiv.org/abs/2408.10069.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation, 2024. URL https://arxiv.org/abs/2311.13385.

Michael C. Fu, Rafael A. Buerba, William D. Long, Daniel J. Blizzard, Andrew W. Lischuk, Andrew H. Haims, and Jonathan N. Grauer. Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions. *The Spine Journal*, 14(10):2442–2448, October 2014. ISSN 1529-9430. doi: 10.1016/j.spinee.2014.03.010. URL http://dx.doi.org/10.1016/j.spinee.2014.03.010.

Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation, 2023. URL https://arxiv.org/abs/2306.13465.

Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, Daguang Xu, and Wenqi Li. Vista3d: Versatile imaging segmentation and annotation model for 3d computed tomography, 2024. URL https://arxiv.org/abs/2406.05285.

R.J. Hemalatha, T.R. Thamizhvani, A. Josephin Arockia Dhivya, Josline Elsa Joseph, Bincy Babu, and R. Chandrasekaran. Active contour based segmentation techniques for medical image analysis. In Robert Koprowski (ed.), *Medical and Biological Image Analysis*, chapter 2. IntechOpen, Rijeka, 2018. doi: 10.5772/intechopen.74576. URL https://doi.org/10.5772/intechopen.74576.

Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of Digital Imaging*, 32(4):582–596, May 2019. ISSN 1618-727X. doi: 10.1007/s10278-019-00227-x. URL http://dx.doi.org/10.1007/s10278-019-00227-x.

Chuanfei Hu, Tianyi Xia, Shenghong Ju, and Xinde Li. When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation, 2023. URL https://arxiv.org/abs/2304.08506.

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

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, December 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z. URL http://dx.doi.org/10.1038/s41592-020-01008-z.

Fabian Isensee, Constantin Ulrich, Tassilo Wald, and Klaus H. Maier-Hein. Extending nnu-net is all you need. In Thomas M. Deserno, Heinz Handels, Andreas Maier, Klaus Maier-Hein, Christoph Palm, and Thomas Tolxdorff (eds.), *Bildverarbeitung für die Medizin 2023*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

Hao Li, Han Liu, Dewei Hu, Jiacheng Wang, and Ipek Oguz. Prism: A promptable and robust interactive segmentation model with visual prompts, 2024. URL https://arxiv.org/abs/2404.15028.

Yanzhen Liu, Sutuke Yibulayimu, Yudi Sang, Gang Zhu, Yu Wang, Chunpeng Zhao, and Xinbao Wu. Pelvic fracture segmentation using a multi-scale distance-weighted neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 312–321. Springer, 2023.

Xiangde Luo, Jia Fu, Yunxin Zhong, Shuolin Liu, Bing Han, Mehdi Astaraki, Simone Bendazzoli, Iuliana Toma-Dasu, Yiwen Ye, Ziyang Chen, et al. Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. *arXiv preprint arXiv:2312.09576*, 2023.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), January 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z. URL http://dx.doi.org/10.1038/s41467-024-44824-z.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, and et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 2024.

Zdravko Marinov, Paul F. Jäger, Jan Egger, Jens Kleesiek, and Rainer Stiefelhagen. Deep interactive segmentation of medical images: A systematic review and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10998–11018, 2024. doi: 10.1109/TPAMI.2024.3452629.

Ahmed W. Moawad, David Fuentes, Ali Morshid, Ahmed M. Khalaf, Mohab M. Elmohr, Abdelrahman Abusaif, John D. Hazle, Ahmed O. Kaseb, Manal Hassan, Armeen Mahvash, Janio Szklaruk, Aliyya Qayyom, and Khaled Elsayes. Multimodality annotated hcc cases with and without advanced imaging segmentation, 2021. URL https://www.cancerimagingarchive.net/collection/hcc-tace-seg/.

Ahmed W. Moawad, Ayahallah A. Ahmed, Mohab ElMohr, Mohamed Eltaher, Mouhammed Amir Habra, Sarah Fisher, Nancy Perrier, Miao Zhang, David Fuentes, and Khaled Elsayes. Voxel-level segmentation of pathologically-proven adrenocortical carcinoma with ki-67 expression (adrenal-acc-ki67-seg), 2023. URL https://www.cancerimagingarchive.net/collection/adrenal-acc-ki67-seg/.

Sovesh Mohapatra, Advait Gosai, and Gottfried Schlaug. Sam vs bet: A comparative study for brain extraction and segmentation of magnetic resonance images using deep learning, 2023. URL https://arxiv.org/abs/2304.04738.

Ali M. Muslim, Syamsiah Mashohor, Gheyath Al Gawwam, Rozi Mahmud, Marsyita binti Hanafi, Osama Alnuaimi, Raad Josephine, and Abdullah Dhaifallah Almutairi. Brain mri dataset of multiple sclerosis with consensus manual lesion segmentation and patient meta information. *Data in Brief*, 2022. doi: https://doi.org/10.1016/j.dib.2022.108139. URL https://www.sciencedirect.com/science/article/pii/S235234092200347X.

Gašper Podobnik, Primož Strojan, Primož Peterlin, Bulat Ibragimov, and Tomaž Vrtovec. Hanseg: The head and neck organ-at-risk ct and mr segmentation dataset. *Medical physics*, 50(3): 1917–1927, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024.

Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R. Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H. Maier-Hein. Sam.md: Zero-shot medical image segmentation capabilities of the segment anything model, 2023. URL https://arxiv.org/abs/2304.05396.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles, 2023. URL https://arxiv.org/abs/2306.00989.

Amber L. Simpson, Jacob Peoples, John M. Creasy, Gabor Fichtinger, Natalie Gangai, Andras Lasso, Krishna Nand Keshava Murthy, Jinru Shia, Michael I. D'Angelica, and Richard K. G. Do. Preoperative ct and survival data for patients undergoing resection of colorectal liver metastases (colorectal-liver-metastases), 2023. URL https://www.cancerimagingarchive.net/collection/colorectal-liver-metastases/.

Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and Klaus H Maier-Hein. Multitalent: A multi-dataset approach to medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 648–658. Springer, 2023.

Kareem Wahid, Cem Dede, Mohamed Naser, and Clifton Fuller. Training dataset for hntsmrg 2024 challenge, 2024. URL https://zenodo.org/doi/10.5281/zenodo.11199559.

Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, Bin Fu, Shaoting Zhang, Junjun He, and Yu Qiao. Sammed3d: Towards general-purpose segmentation models for volumetric medical images, 2024. URL https://arxiv.org/abs/2310.15161.

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

Hallee E. Wong, Marianne Rakic, John Guttag, and Adrian V. Dalca. Scribbleprompt: Fast and flexible interactive segmentation for any biomedical image, 2024. URL https://arxiv.org/abs/2312.07381.

Junde Wu, Jiayuan Zhu, Yueming Jin, and Min Xu. One-prompt to segment all medical images, 2024. URL https://arxiv.org/abs/2305.10300.

Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation, 2023. URL https://arxiv.org/abs/2304.13785.

Binsheng Zhao, Lawrence H Schwartz, Mark G Kris, and Gregory J Riely. Coffee-break lung ct collection with scan images reconstructed at multiple imaging parameters. In *The Cancer Imaging Archive*, 2015. URL https://doi.org/10.7937/k9/tcia.2015.u1x8a5n.

Tao Zhou, Yizhe Zhang, Yi Zhou, Ye Wu, and Chen Gong. Can sam segment polyps?, 2023. URL https://arxiv.org/abs/2304.07583.

A Model specifications

A.1 SAM

SAM is compatible with multiple image encoders, particularly the ViT family from Dosovitskiy et al. (2021). We used the default and best-performing model with ViT-Huge. To ensure high-quality inputs for the model, we performed slice-wise inference by extracting slices from the inplane-plane axis. Each slice was normalized by first clipping values outside the 0.5th and 99.5th percentile of the volume's intensity distribution and then scaling the values to [0, 255]. The image was repeated three times along the channel axis to produce an RGB-like image. Internally, SAM resizes these slices to 1024 pixels for the longest side with the shorter side being padded to 1024 pixels if needed to maintain square dimensions. Finally, the images are normalized using the model's pre-stored mean and standard deviation as suggested by the original implementation. Inference was restricted to slices containing foreground. After prediction, the slices were reassembled into a volume, inverse transformed to the original coordinate system, and metrics were computed in the original image space.

A.2 SAM2

SAM2 supports multiple image encoders, specifically the Hiera family of Ryali et al. (2023). We used the best-performing model, Hiera-L. We clip the intensity values of the volumes based on the 0.5th and 99.5th percentiles, extract each slice along the through-plane, and make the images RGB-like just as with SAM. The images are then rescaled to 1024×1024 pixels and again normalized using the mean and standard deviation provided together with the pretrained weights. Aggregation and inverse transformation are then performed similarly to SAM.

A.3 MedSAM

To apply the model slice-wise, we slice the input volume as with SAM, and then clip each slice based on their 0.5th and 99.5th percentile values. The images are then made RGB-like by repeating thrice along a new channel-dimension, rescaled to 1024×1024 pixels and then normalised to [0, 1]. Aggregation and inverse transformation are performed similarly as with SAM.

A.4 SAM-Med2D

To apply the model slice-by-slice, we slice the input volume as with SAM, and then clip each slice based on their 0.5th and 99.5th percentile values same as with MedSAM. The slices are then made RGB-like and converted to a [0, 255] scale as in SAM's preprocessing. The slices are then standardized using a mean and standard deviation provided along with the model and resized to 256×256 pixels. Aggregation and inverse transformation are performed similarly as with SAM.

A.5 SAM-MED3D

The model is 3D so no slicing is needed. The volume is respaced to $1.5 \times 1.5 \times 1.5$ mm and then clipped based on its 0.5th and 99.5th percentiles. SAMMed3D performs inference on a 128x128x128 crop. The crop is centered around our point prompt if there is only one point prompt passed, and around the centroid of our prompts if multiple points are passed simultaneously. For subsequent refinement steps, the crop remains unchanged. The predicted crop is inserted back in its correct position within the wider coordinate system and then respaced back to the original spacing so that evaluation takes place in the corresponding native image space.

A.6 SAM-MED3D TURBO

SAM-Med3D Turbo is an updated checkpoint for SAMMed-3D and so we perform the same pre- and postprocessing.

A.7 SEGVOL

Intensity values are clipped by its 0.5th and 99.5th percentiles. The mean and standard deviation of the foreground voxels are used for zscore normalization. The values are then rescaled to a [0,1]. Finally, the volume is cropped to its foreground. A first 'zoom-out' inference is performed on this image, followed by a 'zoom-in' sliding window inference. The predicted volume is then transformed back to the original space and compared with the unprocessed ground truth to calculate metrics.

B ADDITIONAL RESULTS

B.1 INITIAL PREDICTION - UNREALISTIC EFFORT

Prompter	Model	Interactions	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Average
1PPS	SAM	1X	0.81	2.66	3.13	2.06	1.42	1.0	7.86	12.47	3.42	3.33	3.82
1PPS	SAM2	1X	1.25	2.74	4.08	3.9	3.07	1.51	9.6	16.35	3.97	3.83	5.03
1PPS	SamMed 2D	1X	10.72	26.3	24.55	28.83	22.76	9.71	30.83	31.24	26.05	20.31	23.13
2±PPS	SAM	2X	0.95	2.79	3.41	3.29	2.12	1.06	8.88	14.31	3.59	3.85	4.43
2±PPS	SAM2	2X	3.39	3.75	7.01	4.89	4.03	1.93	13.17	17.03	4.2	4.97	6.44
2±PPS	SamMed 2D	2X	11.88	28.09	27.62	33.29	24.07	11.25	32.54	31.5	28.27	21.27	24.98
2PPS	SAM	2X	0.88	2.66	3.15	2.56	1.88	1.03	8.1	14.08	3.35	3.7	4.14
2PPS	SAM2	2X	1.6	2.78	4.44	3.9	3.12	1.54	9.82	16.74	3.97	3.93	5.18
2PPS	SamMed 2D	2X	11.63	29.9	28.9	32.88	24.38	11.34	37.84	36.41	32.51	22.39	26.82
3±PPS	SAM	3X	1.02	2.89	3.6	3.61	2.65	1.18	9.66	15.07	3.63	4.08	4.74
3±PPS	SAM2	3X	4.22	4.22	7.78	5.57	4.56	2.08	14.6	17.94	4.36	5.13	7.05
3±PPS	SamMed 2D	3X	12.83	31.8	31.72	36.4	25.85	12.62	40.42	36.91	34.12	23.29	28.6
3PPS	SAM	3X	0.96	2.79	3.18	2.91	2.34	1.18	8.54	14.7	3.44	3.88	4.39
3PPS	SAM2	3X	2.29	3.37	5.64	4.41	3.74	1.79	10.87	17.2	4.09	4.19	5.76
3PPS	SamMed 2D	3X	11.95	31.34	31.8	33.61	24.9	12.15	43.26	40.16	35.62	23.07	28.79
5±PPS	SAM	5X	1.01	2.7	3.98	4.31	2.28	1.31	11.76	15.97	3.87	4.15	5.13
5±PPS	SAM2	5X	7.37	6.15	10.41	6.79	6.57	3.01	23.18	20.41	5.56	6.81	9.63
5±PPS	SamMed 2D	5X	13.95	34.18	35.63	39.39	27.19	14.01	46.49	41.05	37.83	24.5	31.42
5PPS	SAM	5X	1.14	3.37	4.0	3.94	2.96	1.35	10.54	15.5	3.78	4.54	5.11
5PPS	SAM2	5X	4.04	4.67	7.91	5.74	4.5	2.01	15.05	17.77	4.78	4.89	7.14
5PPS	SamMed 2D	5X	12.26	32.17	34.55	33.08	25.24	12.66	50.88	45.0	38.34	23.22	30.74
10±PPS	SAM	10X	2.03	4.77	7.13	8.59	6.32	1.68	18.79	21.29	4.11	9.9	8.46
10±PPS	SAM2	10X	15.61	11.26	15.78	11.5	12.12	7.65	33.5	23.09	7.98	10.01	14.85
10±PPS	SamMed 2D	10X	15.02	35.62	39.67	42.49	28.27	15.95	52.98	46.17	40.97	25.26	34.24
10PPS	SAM	10X	1.3	4.0	5.29	5.3	3.46	1.4	13.45	16.73	4.3	6.66	6.19
10PPS	SAM2	10X	4.61	5.15	7.58	6.12	4.8	2.2	19.8	18.38	5.68	4.79	7.91
10PPS	SamMed 2D	10X	12.03	31.1	36.39	30.38	24.22	13.24	58.07	51.41	38.98	22.1	31.79
Box PS	MedSam	2X	40.63	47.39	55.5	60.4	45.73	46.43	67.75	70.23	48.82	46.13	52.9
Box PS	SAM	2X	13.27	58.73	68.26	63.2	66.22	74.72	70.15	69.46	32.34	62.07	57.84
Box PS	SAM2	2X	70.25	64.18	73.21	73.06	72.07	76.51	73.39	67.9	47.83	65.95	68.44
Box PS	SamMed 2D	2X	28.2	46.08	57.04	64.45	46.44	41.91	62.07	63.59	45.62	41.51	49.69

Table 2: Experimental results simulating unrealistic effort of a clinician prompting each slice of a 3D volume. 'PPS' and 'BPS' represent points per slice or box per slice, respectively. 'X' implies that each interaction is replicated for every slice, multiplying the clinician's effort across the entire volume.

B.2 SINGLE FORWARD PASS - REALISTIC EFFORT 2D

Prompter	Model	Interactions	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Average
3P Inter	SAM	3	0.84	3.6	3.49	2.22	1.67	1.03	8.28	12.67	3.48	4.56	4.18
3P Inter	SAM2	3	1.38	3.67	4.75	4.18	3.41	1.63	10.32	17.01	4.17	5.18	5.57
3P Inter	SamMed 2D	3	11.61	29.08	28.0	34.33	26.0	10.8	36.29	35.8	25.69	23.21	26.08
5P Inter	SAM	5	0.84	3.6	3.5	2.17	1.7	1.02	8.3	12.61	3.64	4.5	4.19
5P Inter	SAM2	5	1.38	3.68	4.73	4.1	3.37	1.63	10.44	16.98	4.4	5.13	5.58
5P Inter	SamMed 2D	5	11.71	29.54	28.46	34.42	26.17	10.88	36.81	35.65	28.27	23.31	26.52
10P Inter	SAM	10	0.84	3.61	3.48	2.21	1.71	1.03	8.26	12.64	3.66	4.49	4.19
10P Inter	SAM2	10	1.38	3.67	4.76	4.14	3.43	1.62	10.39	17.01	4.5	5.1	5.6
10P Inter	SamMed 2D	10	11.74	29.92	28.45	34.71	26.24	10.97	36.67	35.1	29.66	23.33	26.68
5P Prop	SAM	7	1.09	3.49	3.6	1.78	1.7	0.99	7.84	13.7	3.39	4.32	4.19
5P Prop	SAM2	7	3.77	3.75	4.59	3.47	3.13	1.6	8.96	17.52	3.78	4.79	5.54
5P Prop	SamMed 2D	7	10.87	19.33	14.18	17.91	17.09	7.7	27.25	31.84	12.86	13.52	17.26
B Prop	MedSam	4	2.97	20.96	22.75	22.76	23.38	23.81	28.55	29.89	7.16	19.89	20.21
B Prop	SAM	4	0.89	27.98	37.2	37.59	36.38	38.49	43.29	35.47	11.38	30.4	29.91
B Prop	SAM2	4	3.82	33.77	43.31	40.82	41.08	40.88	46.54	37.39	19.37	31.68	33.87
B Prop	SamMed 2D	4	2.82	23.78	30.17	36.12	25.15	22.56	37.98	45.6	19.1	21.0	26.43
3B Inter	MedSam	6	40.14	40.66	46.18	48.16	41.41	40.75	54.42	54.98	30.47	40.63	43.78
3B Inter	SAM	6	13.13	54.04	63.34	58.64	64.26	71.35	65.52	61.78	25.58	59.23	53.69
3B Inter	SAM2	6	69.76	57.59	67.71	69.08	69.06	73.11	68.35	63.24	37.21	63.26	63.84
3B Inter	SamMed 2D	6	27.83	42.22	52.7	60.66	44.76	38.63	55.31	55.81	33.46	38.66	45.0
5B Inter	MedSam	10	40.55	44.91	52.64	57.67	44.75	45.0	64.6	66.88	45.59	44.75	50.73
5B Inter	SAM	10	13.25	56.3	66.81	62.58	65.76	73.63	69.5	68.25	31.33	61.54	56.9
5B Inter	SAM2	10	70.13	61.16	71.53	72.57	71.32	75.52	72.72	67.25	46.01	65.42	67.36
5B Inter	SamMed 2D	10	28.11	44.52	56.04	64.01	46.04	41.24	60.67	62.32	43.44	40.91	48.73
10B Inter	MedSam	20	40.63	46.53	54.89	60.16	45.59	46.15	67.17	69.54	48.69	45.92	52.53
10B Inter	SAM	20	13.27	57.71	67.81	63.17	66.12	74.35	70.02	69.21	32.24	61.98	57.59
10B Inter	SAM2	20	70.25	63.0	72.65	73.05	71.94	76.12	73.19	67.62	47.48	65.83	68.11
10B Inter	SamMed 2D	20	28.19	45.51	56.73	64.4	46.36	41.75	61.89	63.3	45.22	41.4	49.48

Table 3: Experimental results simulating a realistic clinician's effort. 'PPS' and 'PPV' represent points per slice or volume, respectively. 'B Prop' and 'P Prop' denote the introduced box and point propagation schemes, while 'B Inter' and 'P Inter' refer to the introduced box and point interpolation methods.

B.3 SINGLE FORWARD PASS - REALISTIC EFFORT 3D

Prompter	Model	Interactions	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Average
1PPV	SamMed 3D	1	1.92	10.95	21.19	29.12	13.28	16.66	56.06	48.55	42.81	23.63	26.42
1PPV	SamMed 3D Turbo	1	5.03	18.34	30.08	18.38	10.75	34.93	28.51	19.26	39.26	17.06	22.16
1PPV	SegVol	1	8.84	21.28*	31.49	25.49	2.45	32.66	61.77	52.46	29.09	25.85	29.14
1 center PPV	SamMed 3D	1	2.07	12.15	24.11	26.9	15.11	19.64	72.66	53.24	50.4	26.51	30.28
1 center PPV	SamMed 3D Turbo	1	5.18	27.34	46.07	34.46	15.91	46.38	82.98	59.49	63.83	26.98	40.86
1 center PPV	SegVol	1	9.96	24.91*	38.49	31.36	3.17	33.92	71.01	50.67	28.21	30.73	32.24
2 center PPV	SamMed 3D	2	1.87	11.51	23.15	24.45	13.14	18.19	71.11	45.59	40.95	25.36	27.53
2 center PPV	SamMed 3D Turbo	2	5.33	26.62	45.71	33.26	15.75	46.77	84.88	70.5	69.56	27.49	42.59
2 center PPV	SegVol	2	11.2	31.31*	47.51	58.45	11.57	52.36	75.08	52.66	33.4	32.45	40.6
3 center PPV	SamMed 3D	3	1.77	11.29	22.45	23.41	12.19	17.31	70.31	46.24	36.77	24.85	26.66
3 center PPV	SamMed 3D Turbo	3	5.16	26.48	43.81	30.93	15.28	46.56	85.75	68.52	69.67	27.19	41.94
3 center PPV	SegVol	3	11.52	31.1*	50.08	57.51	18.76	53.46	73.04	58.22	45.52	33.95	43.32
5 center PPV	SamMed 3D	5	1.73	10.95	21.86	21.29	11.21	16.04	66.47	43.95	34.78	22.73	25.1
5 center PPV	SamMed 3D Turbo	5	5.09	25.75	42.74	26.85	14.67	46.5	85.92	64.7	70.11	26.6	40.89
5 center PPV	SegVol	5	11.7	31.17*	49.08	52.47	25.4	52.85	61.6	45.66	44.16	33.48	40.76
10 center PPV	SamMed 3D	10	1.72	10.54	20.87	20.78	10.4	14.54	61.01	41.78	27.3	20.16	22.91
10 center PPV	SamMed 3D Turbo	10	5.06	24.9	40.04	21.51	13.1	46.03	86.05	61.5	68.76	25.45	39.24
10 center PPV	SegVol	10	11.69	30.22*	45.24	47.32	26.43	51.68	42.04	26.69	41.88	32.96	35.61
3D Box	SegVol	3	0.55	37.17*	68.11	69.72	63.21	50.13	89.95	79.98	72.13	49.45	58.04

Table 4: Experimental results simulating a realistic clinician's effort. 'center PPV' stands for Point Per Volume, that was sampled from the center of the target object. Sampling from the center performs better than sampling a random point. *SegVol used D2 HanSeg as training data.

B.4 ITERATIVE REFINEMENT 2D

Iterations	Prompter	Model	Interactions	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Average
0	1PPS + 1PPS Refine	SAM	1X/1X	0.81	2.66	3.13	2.06	1.42	1.0	7.86	12.47	3.42	3.33	3.82
1	1PPS + 1PPS Refine	SAM	1X/1X	1.02	2.94	4.46	3.22	1.97	1.13	8.73	13.81	3.5	3.56	4.43
2	1PPS + 1PPS Refine	SAM	1X/1X	1.11	2.97	4.92	3.66	2.5	1.41	10.0	15.55	3.77	3.82	4.97
3	1PPS + 1PPS Refine	SAM	1X/1X	1.18	2.37	5.24	4.41	3.03	1.55	11.06	16.97	4.02	4.0	5.38
4	1PPS + 1PPS Refine	SAM	1X/1X	1.18	2.03	4.64	4.95	3.33	1.62	11.88	17.7	4.16	4.38	5.59
5	1PPS + 1PPS Refine	SAM	1X/1X	1.11	2.2	3.3	5.56	3.37	1.43	12.57	18.25	4.09	4.32	5.62
0	1PPS + 1PPS Refine	SAM2	1X/1X	1.25	2.74	4.08	3.9	3.07	1.51	9.6	16.35	3.97	3.83	5.03
1	1PPS + 1PPS Refine	SAM2	1X/1X	2.96	4.99	8.69	6.2	5.64	3.32	17.94	24.31	5.49	5.78	8.53
2	1PPS + 1PPS Refine	SAM2	1X/1X	5.25	6.83	11.68	8.03	7.64	5.04	24.56	29.52	7.39	7.83	11.38
3	1PPS + 1PPS Refine	SAM2	1X/1X	7.53	8.8	15.02	10.01	9.52	6.87	29.11	33.53	9.06	9.59	13.9
4	1PPS + 1PPS Refine	SAM2	1X/1X	9.81	10.51	18.31	11.99	11.09	8.24	33.22	48.49	10.45	11.54	17.37
5	1PPS + 1PPS Refine	SAM2	1X/1X	12.14	12.15	21.33	13.52	12.56	9.21	37.04	51.38	11.72	13.5	19.45
0	1PPS + 1PPS Refine	SamMed 2D	1X/1X	10.72	26.3	24.55	28.83	22.76	9.71	30.83	31.24	26.05	20.31	23.13
1	1PPS + 1PPS Refine	SamMed 2D	1X/1X	12.44	29.97	28.54	31.35	24.3	11.78	39.25	37.54	31.41	21.45	26.8
2	1PPS + 1PPS Refine	SamMed 2D	1X/1X	12.49	31.19	30.25	32.28	24.65	12.41	43.74	40.46	33.49	22.07	28.3
3	1PPS + 1PPS Refine	SamMed 2D	1X/1X	12.59	31.93	31.49	32.85	24.98	12.88	46.69	42.32	34.71	22.6	29.31
4	1PPS + 1PPS Refine	SamMed 2D	1X/1X	12.78	32.51	32.59	33.31	25.33	13.31	48.82	43.68	35.59	23.05	30.1
5	1PPS + 1PPS Refine	SamMed 2D	1X/1X	12.93	32.98	33.67	33.74	25.76	13.67	50.46	44.81	36.29	23.49	30.78
0	1PPS + Scribble Refine	SAM	1/3	0.81	2.66	3.13	2.06	1.42	1.0	7.86	12.47	3.42	3.33	3.82
1	1PPS + Scribble Refine	SAM	1/3	0.89	2.65	4.53	3.24	2.11	1.18	7.48	11.65	3.55	3.68	4.1
2	1PPS + Scribble Refine	SAM	1/3	0.93	2.76	4.45	3.54	2.35	1.37	8.01	11.24	3.52	3.41	4.16
3	1PPS + Scribble Refine	SAM	1/3	0.96	2.12	3.5	3.49	2.09	1.4	8.01	10.6	3.4	3.15	3.87
4	1PPS + Scribble Refine	SAM	1/3	0.92	1.94	2.08	3.43	1.94	1.4	8.06	9.84	3.27	3.01	3.59
5	1PPS + Scribble Refine	SAM	1/3	0.88	1.69	1.4	3.43	1.87	1.14	7.66	9.54	3.14	2.56	3.33
0	1PPS + Scribble Refine	SAM2	1/3	1.25	2.74	4.08	3.9	3.07	1.51	9.6	16.35	3.97	3.83	5.03
1	1PPS + Scribble Refine	SAM2	1/3	1.74	3.88	8.13	5.39	5.97	3.01	16.14	18.28	4.66	5.07	7.23
2	1PPS + Scribble Refine	SAM2	1/3	2.18	4.42	10.05	5.84	7.57	4.14	19.28	19.15	4.86	5.75	8.32
3	1PPS + Scribble Refine	SAM2	1/3	2.52	4.8	10.93	6.16	8.2	5.06	21.87	25.97	4.9	6.17	9.66
4	1PPS + Scribble Refine	SAM2	1/3	2.83	5.03	12.37	6.26	8.75	5.88	24.46	27.84	4.86	6.45	10.47
5	1PPS + Scribble Refine	SAM2	1/3	3.1	5.3	13.38	6.26	9.08	6.93	25.81	30.02	4.86	6.85	11.16
0	1PPS + Scribble Refine	SamMed 2D	1/3	10.72	26.3	24.55	28.83	22.76	9.71	30.83	31.24	26.05	20.31	23.13
1	1PPS + Scribble Refine	SamMed 2D	1/3	11.8	28.84	26.82	29.72	23.07	11.32	37.22	36.7	29.87	20.7	25.61
2	1PPS + Scribble Refine	SamMed 2D	1/3	11.71	29.29	27.88	29.5	22.7	11.63	40.64	38.77	31.61	20.65	26.44
3	1PPS + Scribble Refine	SamMed 2D	1/3	11.55	29.52	28.26	29.15	22.67	11.68	42.57	39.92	32.62	20.68	26.86
4	1PPS + Scribble Refine	SamMed 2D	1/3	11.49	29.62	28.33	29.19	22.6	11.76	43.94	40.71	33.25	20.74	27.16
5	1PPS + Scribble Refine	SamMed 2D	1/3	11.46	29.75	28.72	28.87	22.52	11.88	44.93	41.32	33.68	20.83	27.4
0	3B Inter + Scribble Refine	SAM	6/3	13.13	54.04	63.34	58.64	64.26	71.35	65.52	61.78	25.58	59.23	53.69
1	3B Inter + Scribble Refine	SAM	6/3	0.99	3.66	3.24	3.53	2.33	6.14	7.78	11.88	4.68	6.43	5.07
2	3B Inter + Scribble Refine	SAM	6/3	0.85	1.95	1.28	3.17	9.04	5.28	7.49	14.4	4.14	7.43	5.51
3	3B Inter + Scribble Refine	SAM	6/3	0.8	4.11	2.95	2.19	3.13	3.36	7.14	9.85	4.02	4.67	4.22
4	3B Inter + Scribble Refine	SAM	6/3	0.71	2.45	2.01	2.21	2.28	1.89	6.08	9.09	3.22	3.57	3.35
5	3B Inter + Scribble Refine	SAM	6/3	0.76	1.99	2.46	2.51	2.16	1.42	6.15	9.34	3.19	3.47	3.34
0	3B Inter + Scribble Refine	SAM2	6/3	69.78	57.59	67.71	69.08	69.06	73.11	68.35	63.24	37.21	63.27	63.84
1	3B Inter + Scribble Refine	SAM2	6/3	10.03	20.17	25.37	19.13	23.65	18.12	25.37	27.66	11.34	15.93	19.68
2	3B Inter + Scribble Refine	SAM2	6/3	5.07	6.98	10.53	7.31	8.35	3.99	20.7	19.58	5.61	6.41	9.45
3	3B Inter + Scribble Refine	SAM2	6/3	4.23	4.43	7.8	6.93	5.64	3.72	14.43	16.35	4.89	5.34	7.38
4	3B Inter + Scribble Refine	SAM2	6/3	3.99	3.84	6.96	7.22	4.86	3.53	13.37	15.02	4.56	4.9	6.83
5	3B Inter + Scribble Refine	SAM2	6/3	3.41	3.7	6.75	5.7	3.99	2.79	12.49	13.59	4.3	4.47	6.12
0	3B Inter + Scribble Refine	SamMed 2D	6/3	27.83	42.22	52.7	60.66	44.76	38.63	55.31	55.81	33.46	38.66	45.0
1	3B Inter + Scribble Refine	SamMed 2D	6/3	28.4	44.72	57.53	63.33	45.28	41.42	61.61	59.73	40.44	38.8	48.13
2	3B Inter + Scribble Refine	SamMed 2D	6/3	24.29	42.41	53.02	59.24	40.27	35.69	62.86	58.85	41.23	34.93	45.28
3	3B Inter + Scribble Refine	SamMed 2D	6/3	20.91	39.69	47.48	52.73	35.55	28.2	62.08	54.89	40.02	31.59	41.31
4	3B Inter + Scribble Refine	SamMed 2D	6/3	18.76	37.29	41.46	46.5	31.73	23.98	58.46	51.86	38.07	28.74	37.69
5	3B Inter + Scribble Refine	SamMed 2D	6/3	17.07	35.06	37.39	41.62	29.1	20.8	53.9	49.27	36.27	26.66	34.71
0	3P Inter + Scribble Refine	SAM	5/3	0.84	3.6	3.5	2.17	1.7	1.02	8.3	12.61	3.64	4.5	4.19
1	3P Inter + Scribble Refine	SAM	5/3	0.89	3.29	4.59	3.58	2.41	1.21	7.96	11.68	3.73	4.39	4.37
2	3P Inter + Scribble Refine	SAM	5/3	0.98	3.19	5.18	4.21	2.65	1.44	8.64	11.54	3.68	4.24	4.58
3	3P Inter + Scribble Refine	SAM	5/3	1.03	2.5	3.61	4.14	2.52	1.5	9.12	10.95	3.64	3.71	4.27
4	3P Inter + Scribble Refine	SAM	5/3	1.03	2.04	2.23	4.14	2.31	1.48	8.91	10.4	3.48	3.32	3.93
5	3P Inter + Scribble Refine	SAM	5/3	0.99	1.89	1.5	4.17	1.87	1.32	8.16	9.75	3.33	3.02	3.6
0	3P Inter + Scribble Refine	SAM2	5/3	1.38	3.68	4.73	4.1	3.37	1.63	10.44	16.98	4.4	5.13	5.58
1	3P Inter + Scribble Refine	SAM2	5/3	1.87	5.18	8.77	6.04	6.8	3.47	16.57	19.63	5.09	6.31	7.97
2	3P Inter + Scribble Refine	SAM2	5/3	2.37	5.83	11.33	7.01	9.24	4.62	22.39	21.17	5.41	7.22	9.66
3	3P Inter + Scribble Refine	SAM2	5/3	2.84	6.27	13.65	7.62	10.71	5.9	23.75	27.73	5.48	7.65	11.16
4	3P Inter + Scribble Refine	SAM2	5/3	3.33	6.52	14.69	7.83	11.49	6.96	27.72	28.86	5.52	8.05	12.1
5	3P Inter + Scribble Refine	SAM2	5/3	3.5	6.7	15.79	7.97	12.18	7.65	30.01	30.82	5.56	8.47	12.86
0	3P Inter + Scribble Refine	SamMed 2D	5/3	11.71	29.54	28.46	34.42	26.17	10.88	36.81	35.65	28.27	23.31	26.52
1	3P Inter + Scribble Refine	SamMed 2D	5/3	12.72	31.26	30.27	34.05	26.02	12.06	40.92	38.13	31.14	23.12	27.97
2	3P Inter + Scribble Refine	SamMed 2D	5/3	12.46	31.66	30.58	33.13	25.57	12.19	42.79	39.25	32.55	22.88	28.31
3	3P Inter + Scribble Refine	SamMed 2D	5/3	12.25	31.79	30.53	32.57	25.24	12.28	43.89	39.98	33.35	22.73	28.46
4	3P Inter + Scribble Refine	SamMed 2D	5/3	12.13	31.83	30.42	32.12	25.05	12.33	44.89	40.52	33.84	22.63	28.58
5	3P Inter + Scribble Refine	SamMed 2D	5/3	12.06	31.86	30.42	31.66	24.9	12.41	45.85	40.97	34.22	22.61	28.7

Table 5: Interactive refinement results for 2D models across 5 iterations. The initial prediction is made either using a single point per slice or one of our proposed prompting schemes. Omitting the previous point during refinement led to a drop in performance, which explains the drop in performance for the 3 Box interpolation after the initial prompt.The unrealistic slice-wise refinement (1 interaction per slice) is only slightly better than our proposed scribble refinement method (3 interactions).

B.5 ITERATIVE REFINEMENT 3D

Iterations	Prompter	Model	Interactions	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Average
0	1 center PPV + 1 PPV Refine + prev point	SamMed 3D	1/1	2.03	12.15	24.11	26.9	15.09	19.64	72.66	53.03	50.4	26.51	30.25
1	1 center PPV + 1 PPV Refine + prev point	SamMed 3D	1/1	5.89	14.07	33.03	35.67	14.42	24.93	64.15	72.35	42.64	24.86	33.2
2	1 center PPV + 1 PPV Refine + prev point	SamMed 3D	1/1	5.67	13.49	29.52	28.92	14.34	21.4	64.29	62.44	42.03	25.16	30.73
3	1 center PPV + 1 PPV Refine + prev point	SamMed 3D	1/1	5.57	13.64	31.57	35.51	14.21	22.82	63.35	65.94	41.12	24.79	31.85
4	1 center PPV + 1 PPV Refine + prev point	SamMed 3D	1/1	5.81	13.51	31.44	31.08	14.11	21.38	65.28	63.38	40.68	24.82	31.15
5	1 center PPV + 1 PPV Refine + prev point	SamMed 3D	1/1	5.46	13.56	32.19	33.49	14.18	24.28	62.12	65.2	40.41	25.12	31.6
0	1 center PPV + 1 PPV Refine + prev point	SamMed 3D Turbo	1/1	5.12	27.34	46.07	34.46	15.9	46.38	82.98	59.36	63.83	26.98	40.84
1	1 center PPV + 1 PPV Refine + prev point	SamMed 3D Turbo	1/1	2.81	21.83	27.6	17.69	8.25	30.88	79.91	72.82	61.07	21.37	34.42
2	1 center PPV + 1 PPV Refine + prev point	SamMed 3D Turbo	1/1	3.57	22.72	35.57	22.05	8.89	36.31	77.77	72.2	62.7	23.37	36.52
3	1 center PPV + 1 PPV Refine + prev point	SamMed 3D Turbo	1/1	3.25	22.22	32.87	17.88	8.74	33.27	77.73	66.93	61.51	22.17	34.66
4	1 center PPV + 1 PPV Refine + prev point	SamMed 3D Turbo	1/1	3.6	22.93	34.4	16.26	8.78	34.84	80.01	72.09	62.35	23.35	35.86
5	1 center PPV + 1 PPV Refine + prev point	SamMed 3D Turbo	1/1	3.09	21.72	32.32	14.83	7.66	31.79	76.55	70.69	62.1	22.19	34.3
0	1 center PPV + 1 PPV Refine	SamMed 3D	1/1	2.03	12.15	24.11	26.9	15.09	19.64	72.66	53.04	50.4	26.51	30.25
1	1 center PPV + 1 PPV Refine	SamMed 3D	1/1	3.16	13.08	27.85	36.99	14.07	23.65	72.64	78.47	49.24	26.72	34.59
2	1 center PPV + 1 PPV Refine	SamMed 3D	1/1	4.12	12.99	28.79	36.47	12.5	24.22	72.88	82.11	48.78	26.71	34.96
3	1 center PPV + 1 PPV Refine	SamMed 3D	1/1	4.56	12.97	28.68	35.59	11.36	24.71	72.23	82.5	49.52	26.54	34.87
4	1 center PPV + 1 PPV Refine	SamMed 3D	1/1	4.9	13.11	27.59	36.86	10.75	24.55	72.48	82.87	49.31	27.01	34.94
5	1 center PPV + 1 PPV Refine	SamMed 3D	1/1	4.94	13.04	27.36	36.85	10.35	25.03	73.23	82.92	49.42	26.95	35.01
0	1 center PPV + 1 PPV Refine	SamMed 3D Turbo	1/1	5.12	27.34	46.07	34.46	15.9	46.38	82.98	59.37	63.83	26.98	40.84
1	1 center PPV + 1 PPV Refine	SamMed 3D Turbo	1/1	5.65	28.81	48.08	38.1	16.41	50.17	86.5	73.12	67.78	28.59	44.32
2	1 center PPV + 1 PPV Refine	SamMed 3D Turbo	1/1	5.82	29.45	48.88	43.04	16.94	51.87	87.34	79.33	69.56	29.79	46.2
3	1 center PPV + 1 PPV Refine	SamMed 3D Turbo	1/1	5.86	30.04	49.11	46.83	17.56	53.0	87.71	80.79	70.36	30.81	47.21
4	1 center PPV + 1 PPV Refine	SamMed 3D Turbo	1/1	5.97	30.56	49.35	45.66	18.3	54.12	87.78	82.91	71.25	31.6	47.75
5	1 center PPV + 1 PPV Refine	SamMed 3D Turbo	1/1	6.17	30.92	50.24	49.17	19.24	54.74	88.15	84.43	71.94	32.13	48.71
0	1 center PPV + Scribble Refine	SamMed 3D	1/3	2.03	12.15	24.11	26.9	15.09	19.64	72.66	53.04	50.4	26.51	30.25
1	1 center PPV + Scribble Refine	SamMed 3D	1/3	3.31	12.58	25.65	32.52	13.57	23.05	71.32	68.61	47.72	26.04	32.44
2	1 center PPV + Scribble Refine	SamMed 3D	1/3	4.03	12.93	25.8	34.42	11.66	24.54	72.38	78.9	47.89	26.2	33.88
3	1 center PPV + Scribble Refine	SamMed 3D	1/3	4.42	12.93	26.36	34.38	10.89	24.88	73.54	84.12	48.88	25.95	34.64
4	1 center PPV + Scribble Refine	SamMed 3D	1/3	4.37	13.05	26.78	35.61	10.29	25.33	73.91	85.34	49.14	25.81	34.96
5	1 center PPV + Scribble Refine	SamMed 3D	1/3	4.56	13.23	27.32	36.3	9.85	25.76	73.7	86.4	49.55	25.87	35.25
0	1 center PPV + Scribble Refine	SamMed 3D Turbo	1/3	5.12	27.34	46.07	34.46	15.9	46.38	82.98	59.37	63.83	26.98	40.84
1	1 center PPV + Scribble Refine	SamMed 3D Turbo	1/3	5.36	27.71	47.84	37.92	15.93	48.96	86.18	71.33	67.78	27.93	43.69
2	1 center PPV + Scribble Refine	SamMed 3D Turbo	1/3	4.84	28.19	48.07	40.22	16.6	50.74	87.2	76.94	69.93	29.01	45.17
3	1 center PPV + Scribble Refine	SamMed 3D Turbo	1/3	4.34	28.75	47.42	42.91	17.08	52.24	87.71	78.53	70.86	30.08	45.99
4	1 center PPV + Scribble Refine	SamMed 3D Turbo	1/3	4.22	29.41	47.72	43.91	17.65	53.4	88.06	80.09	71.9	31.06	46.74
5	1 center PPV + Scribble Refine	SamMed 3D Turbo	1/3	4.3	29.84	48.23	45.84	17.89	54.62	88.24	80.88	72.54	31.66	47.4

Table 6: Interactive refinement results for 3D models over 5 iterations. The initial interaction always starts from a central point of the target object, and refinement is performed either by randomly sampling positive or negative points (1 interaction) or by selecting a point using the proposed scribble refinement method. Scribble drawing is counted as three interactions. In contrast to 2D models, including the previous point prompt did not improve the performance.