UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

From point to mask: interactive segmentation for 3D ultrasound

Zoe Reinke, Tom J. Blöcker, Guillaume Landry, Thomas Wendler

Angaben zur Veröffentlichung / Publication details:

Reinke, Zoe, Tom J. Blöcker, Guillaume Landry, and Thomas Wendler. 2026. "From point to mask: interactive segmentation for 3D ultrasound." *Current Directions in Biomedical Engineering* 12 (1): 436–39. https://doi.org/10.1515/cdbme-2026-0211.

Nutzungsbedingungen / Terms of use:

CC BY 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

Zoe Reinke*, Tom Julius Blöcker, Guillaume Landry, and Thomas Wendler

From Point to Mask: Interactive Segmentation for 3D Ultrasound

https://doi.org/10.1515/cdbme-2026-0211

Abstract: Segmentation is an important tool in medical imaging but relies on time-consuming manual annotation. Artificial intelligence offers a promising route to support medical professionals, with many models available targeting different modalities and dimensionalities. Most existing work focuses either on 2D images or on three-dimensional CT and MRI, leaving 3D ultrasound (3DUS) largely unexplored. This work evaluates three promptable segmentation models not specifically trained for 3DUS across multiple prompt types and four openly available datasets, to assess their feasibility for this challenging modality.

Keywords: Interactive segmentation, three-dimensional ultrasound, foundation models, promptable models

1 Introduction

Recent advances in artificial intelligence have produced a growing number of openly available image segmentation models trained on large datasets. Training a task-specific model requires extensive annotated data, which is often unavailable, making the reuse of existing general-purpose models particularly attractive. Prior work has begun evaluating such models across broad modality spectra [6] or for specific clinical tasks [1]. To our knowledge, no systematic comparison exists for 3D ultrasound (3DUS), reflecting the general underrepresentation of ultrasound, especially 3DUS, in 3D medical training datasets relative to MRI, CT, and PET.

***Corresponding author: Zoe Reinke,** Dept. of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Augsburg, Stenglinstraße 2, Augsburg, Germany, e-mail: zoe.reinke@med.uni-augsburg.de
Tom Julius Blöcker, Guillaume Landry, Dept. of Radiation Oncology, Medical Center of the University of Munich, Germany
Guillaume Landry, Thomas Wendler, Bavarian Center for Cancer Research (BZKF), Germany
Thomas Wendler, Dept. of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Augsburg
Thomas Wendler, Zoe Reinke, Digital Medicine, University Hospital Augsburg, Augsburg, Germany
Thomas Wendler, Chair for Computer Aided Medical Procedures and Augmented Reality, Technical University of Munich, Germany
Thomas Wendler, Center for Advanced Analytics and Predictive Sciences, University of Augsburg, Germany

Two architectural families are represented in this study. SAM2 [9] serves as a general-purpose foundation model, and its medically fine-tuned variant MedSAM2 [7] represents domain adaptation to diverse medical imaging data. Additionally, nnInteractive [4], a U-Net-based model trained exclusively on medical volumetric data, was included as a model designed specifically for 3D interactive segmentation in the broader domain. These models were run on a multitude of prompt and dataset combinations, to reflect the broad spectrum of variability in the clinical setting, and to evaluate whether higher manual effort results in better segmentations.

2 Experiment Setup

2.1 Models

SAM2 was evaluated using the *SAM2.1-large* checkpoint. For MedSAM2, a generalized checkpoint and a checkpoint fine-tuned for cardiac ultrasound were compared on a data subset; the ultrasound-specific checkpoint showed slightly higher performance and was used throughout. For nnInteractive, version 1.1.3 was used.

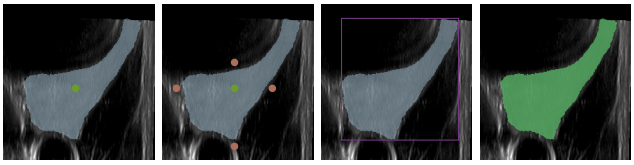
2.2 Data

Ultrasound appearance varies substantially across anatomies, with some structures presenting weak delineation from surrounding tissue. To evaluate across diverse anatomies, publicly available 3DUS datasets with ground truth segmentation masks were sought. Due to the limited availability of such data, only four were identified:

- **Leg-3D-US**: combined scans of three leg muscles [3]
- **SegThy**: thyroid gland, carotid artery, and jugular vein [5]
- **TRUSTED**: kidney [8]
- **UterUS**: uterus [2]

All datasets included ground truth annotation masks. NIfTI volumes were sliced along the ultrasound sweep axis to produce image stacks representing the acquisition sequence. nnInteractive accepts NIfTI volumes directly and was used accordingly.

Open Access. © 2026 The Author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 License.



(a) Single-point **(b)** Multi-point **(c)** Bounding box **(d)** Mask

Fig. 1: Prompt types used in this study, illustrated for the thyroid of a file from the SegThy dataset.

2.3 Prompts

Four prompt types were evaluated, reflecting different levels of manual effort:

- **Single-point:** one positive point at the object centroid
- **Multi-point:** one positive point at the object centroid and four negative points placed outside the mask along cardinal directions
- **Bounding box:** a tight bounding box around the object
- **Mask:** the ground truth segmentation mask

All prompts were derived from the ground truth and applied at the central slice of the sweep axis for each labeled object, which was assumed to provide maximal structural information for propagation. For the SAM-based models, segmentation was propagated forward and backward from this slice, and the results were combined.

3 Results

Three models were evaluated across four prompt types, yielding twelve experiment configurations per file. Ten files were selected from each dataset to balance dataset sizes, giving 40 files in total. Files were chosen to capture variability in each dataset, for example for different patients, body sides and data centers. Segmentation performance was assessed using the Dice Similarity Coefficient (DSC), measuring volumetric overlap between prediction and ground truth, and the 95th percentile Hausdorff Distance (HD95), quantifying boundary accuracy while limiting sensitivity to outliers. Since data was not normally distributed (Shapiro-Wilk, $p < 0.05$ for all but one group), statistical significance was tested using Mann-Whitney U tests.

3.1 Prompt comparison

A clear prompt-dependent trend is visible in Figure 2 and Table 1. Mask prompts achieved the highest mean DSC across all models (0.802–0.817), followed closely by bounding box prompts (0.731–0.776). Single-point and multi-point prompts resulted in substantially lower performance, consistent with findings in other modalities [1, 6].

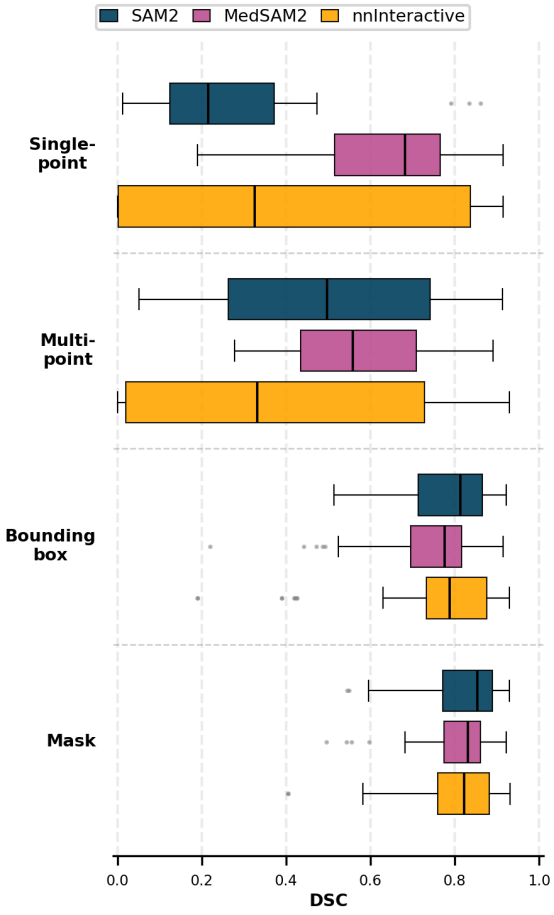


Fig. 2: DSC distributions for all prompt types and models. The line inside each box indicates the median; boxes show the interquartile range.

The relationship between single- and multi-point prompts varied across models. For SAM2, multi-point prompts significantly outperformed single-point prompts ($p < 0.001$), indicating sensitivity to negative point guidance. For MedSAM2 and nnInteractive, no statistically significant difference was found ($p = 0.082$ and $p = 0.610$ respectively), and for MedSAM2 the single-point mean DSC (0.625) even slightly exceeded the multi-point mean (0.568), suggesting the model extracts sufficient context from a single centroid prompt and may be sensitive to suboptimal negative point placement.

3.2 Model comparison

Table 1 summarises DSC per model and prompt type. For mask and bounding box prompts, all three models performed

Tab. 1: DSC per prompt type and model (mean ± standard deviation). Best value per prompt type in **bold**.

Prompt	Model	DSC
Single-point	SAM2	0.275 ± 0.201
	MedSAM2	$\mathbf{0.625 \pm 0.186}$
	nnInteractive	0.387 ± 0.374
Multi-point	SAM2	0.500 ± 0.263
	MedSAM2	$\mathbf{0.568 \pm 0.160}$
	nnInteractive	0.410 ± 0.347
Bounding box	SAM2	$\mathbf{0.776 \pm 0.110}$
	MedSAM2	0.731 ± 0.146
	nnInteractive	0.763 ± 0.158
Mask	SAM2	$\mathbf{0.817 \pm 0.098}$
	MedSAM2	0.802 ± 0.100
	nnInteractive	0.804 ± 0.102

comparably, with no statistically significant difference between any model pair (all $p > 0.12$), and SAM2 achieving the highest mean DSC for box and mask prompts (0.776 and 0.817 respectively).

MedSAM2 showed the best mean performance for single-point and multi-point prompts (0.625 and 0.568), while nnInteractive performed substantially worse (0.387 and 0.410), frequently due to extreme undersegmentation caused by the model's AutoZoom mechanism failing to expand its segmentation patch for heterogeneous structures or volumes with acquisition artefacts. This leads to a bimodal outcome: either a good segmentation or near-complete failure, visible as a wide interquartile range in Figure 2 for both point-based prompt types, where for multi-point prompts the median lies substantially below the mean values reported in Table 1.

SAM2 and nnInteractive performed comparably for both point-based prompt types, with no statistically significant difference between them ($p = 0.861$ and $p = 0.133$ respectively). For single-point prompts, MedSAM2 significantly outperformed both SAM2 ($p < 0.001$) and nnInteractive ($p = 0.024$). For multi-point prompts, MedSAM2 significantly outperformed nnInteractive ($p = 0.026$) but did not significantly differ from SAM2 ($p = 0.279$).

Overall, MedSAM2 dominated for point prompts, whereas SAM2 achieved the highest scores for detailed prompts, followed by nnInteractive. This shows a clear trend: fine-tuned MedSAM2 works much better with little information, whereas the other models perform best only for more detailed prompts. The stronger performance of SAM2 over nnInteractive for detailed prompts may reflect that prompts were applied in 2D, disadvantaging nnInteractive which is designed for 3D volumetric prompting.

3.3 Dataset differences

Because anatomical scale differs greatly across datasets, HD95 values are not meaningful when pooled and are reported per dataset in Table 2. These values are aggregated across all models and prompt types to capture full dataset variability. Leg-3D-US showed the highest HD95 values (94.97 ± 88.14 voxels), attributable to the large scan volume resulting from combining multiple sweeps to cover the full leg length. Smaller structures such as UterUS achieved substantially lower HD95 values (14.97 ± 20.27 voxels). DSC values were more consistent across datasets (0.568–0.746), indicating that volumetric overlap was generally maintained despite large differences in boundary accuracy.

Tab. 2: DSC and HD95 per dataset (mean ± standard deviation), aggregated across all models and prompts. HD95 reported in voxels. Best value per metric in **bold**.

Dataset	DSC	HD95
Leg-3D-US	0.568 ± 0.265	94.97 ± 88.14
SegThy	$\mathbf{0.746 \pm 0.233}$	40.75 ± 59.56
TRUSTED	0.628 ± 0.319	57.11 ± 64.24
UterUS	0.571 ± 0.256	$\mathbf{14.97 \pm 20.27}$

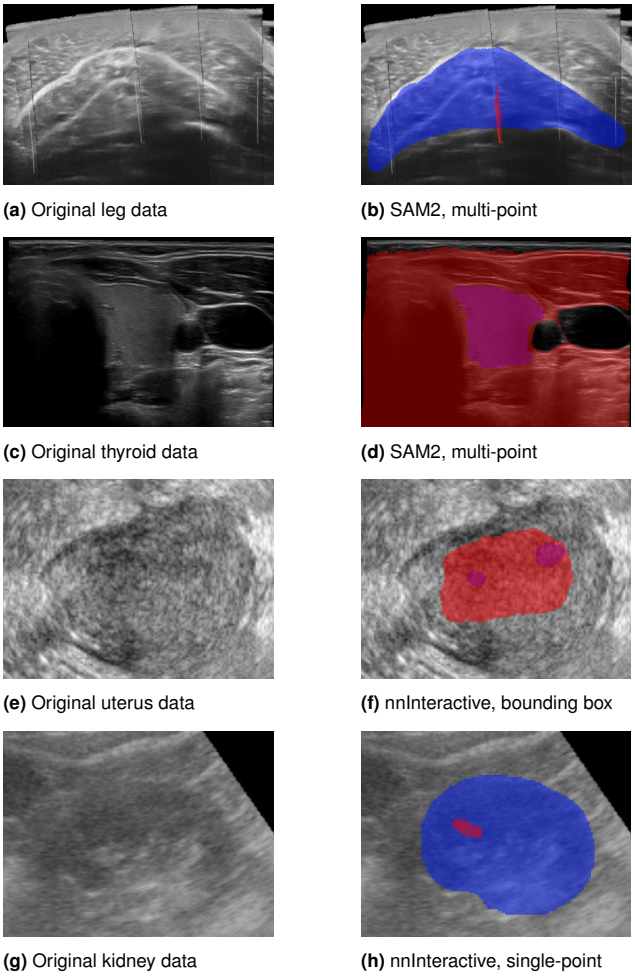
3.4 Failure cases

Figure 3 shows representative failure cases. Reasons for failure include image artefacts (aligned scans overlap and create edges for leg muscle data), hard to delineate borders, and heterogeneity of tissue leading to over- or under-segmentation. Point-based prompts were most commonly associated with failures, as they provide insufficient spatial constraints to overcome these challenges.

3.5 Discussion

Single-point and multi-point prompts require the least manual effort but were consistently outperformed by bounding box and mask prompts as expected. Differences between models were relatively low for bounding box and mask prompts, showing that all models offer a good base for semi-automatic segmentation in a clinical workflow.

The results demonstrated some variability between files and configurations, leading to outliers or failures. Visual inspection revealed two failure modes. For data with unclear boundaries, a tendency towards oversegmentation was observed, especially for SAM-based methods. For heterogeneous



(a) Original leg data **(b)** SAM2, multi-point

(c) Original thyroid data **(d)** SAM2, multi-point

(e) Original uterus data **(f)** nnInteractive, bounding box

(g) Original kidney data **(h)** nnInteractive, single-point

Fig. 3: Failure cases from all datasets for SAM2 and nnInteractive. Ground truth is highlighted in blue, prediction in red and overlap in pink.

tissue, smaller homogeneous sub-regions or artefacts were incorrectly identified as anatomical borders, leading to under-segmentation.

4 Conclusion

This study demonstrates that existing promptable segmentation models can be applied to 3DUS with clinically relevant performance, particularly for bounding box and mask prompts. The choice of model and prompt type influences the outcome, and the cost-benefit trade-off between prompt effort and segmentation quality should guide clinical adoption.

Prompts in this study were placed at the centre of the ground truth mask on the middle sweep slice, which may not reflect clinical practice where prompts are placed from anatomical knowledge. Future work should integrate prompt variation and explore multi-axis prompting for SAM-based models to fully exploit the three-dimensionality of the data.

Author Statement

Research funding: This research was partially funded by the Intramural Research Funding Grants “NeUSBrain” and “3DUS-ICAS” of the Faculty of Medicine, University of Augsburg, as well as by the Bavarian Ministry of Economic Affairs, Regional Development and Energy (StMWi) under grant number LSM-2403-0017. **Conflict of interest**: Authors state no conflict of interest. **Informed consent**: Publicly available datasets were used, where informed consent information is not given. **Ethical approval**: Publicly available datasets were used, where ethics approval by the responsible institutional review board is claimed by the authors of the respective studies that created the data.

References

[1] Blöcker JT, Delopoulos N, Palacios MA, et al. GTV segmentation in MRI guided radiotherapy with promptable foundation models. Phys Med Biol. 2025;71(1). doi:10.1088/1361-6560/ae2db9

[2] Boneš E, Gergolet M, Bohak C, Lesar Ž, Marolt M. Automatic segmentation and alignment of uterine shapes from 3D ultrasound data. Comput Biol Med. 2024;178:108794. doi:10.1016/j.compbiomed.2024.108794

[3] Duque VG, Marquardt A, Velikova Y, et al. Ultrasound segmentation analysis via distinct and completed anatomical borders. Int J Comput Assist Radiol Surg. 2024;19(7):1419-1427. doi:10.1007/s11548-024-03170-7

[4] Isensee F, Rokuss M, Krämer L, et al. nnInteractive: Redefining 3D Promptable Segmentation. arXiv preprint arXiv:2503.08373. 2025.

[5] Krönke M, Eilers C, Dimova D, et al. Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. PLoS One. 2022;17(7):e0268550. doi:10.1371/journal.pone.0268550

[6] Lurz D, Neubig L, Kopp M, Kist A. Foundation Models in Medical Image Segmentation. In: Handels H, et al., editors. Bildverarbeitung für die Medizin 2026. BVM 2026. Informatik aktuell. Wiesbaden: Springer Vieweg; 2026. doi:10.1007/978-3-658-51100-5_8

[7] Ma J, Yang Z, Kim S, et al. MedSAM2: Segment Anything in 3D Medical Images and Videos. arXiv preprint arXiv:2504.03600. 2025.

[8] Ndzimbong W, Fourniol C, Themyr L, et al. TRUSTED: The paired 3D transabdominal ultrasound and CT human data for kidney segmentation and registration research. Sci Data. 2025;12(1):615. doi:10.1038/s41597-025-04467-1

[9] Ravi N, Gabeur V, Hu YT, et al. SAM 2: Segment Anything in Images and Videos. arXiv preprint arXiv:2408.00714. 2024.