Physics in Medicine & Biology

PAPER • OPEN ACCESS

GTV segmentation in MRI guided radiotherapy with promptable foundation models

To cite this article: Tom Julius Blöcker *et al* 2026 *Phys. Med. Biol.* **71** 015006

Physics in Medicine & Biology

IPEM
Institute of Physics and Engineering in Medicine

Check for updates

OPEN ACCESS

RECEIVED
12 August 2025

REVISED
9 December 2025

ACCEPTED FOR PUBLICATION
16 December 2025

PUBLISHED
29 December 2025

PAPER

GTV segmentation in MRI guided radiotherapy with promptable foundation models

Tom Julius Blöcker[1], Nikolaos Delopoulos[1], Miguel A Palacios[2], Sebastian Klüter[3], Juliane Hörner-Rieber[3,4], Carolin Rippke[3], Lorenzo Placidi[5], Luca Boldrini[6,7], Vincenzo Frascino[7], Nicolaus Andratschke[8], Michael Baumgartl[8], Riccardo Dal Bello[8], Sebastian N Marschner[1], Claus Belka[1,9,10], Stefanie Corradini[1,11], Denis Dudas[1], Marco Riboldi[12], Christopher Kurz[1] and Guillaume Landry[1,10,*]

[1] Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany
[2] Amsterdam UMC, Vrije Universiteit Medical Centre, Dept. of Radiation Oncology, de Boelelaan 1117, 1081 HV Amsterdam, The Netherlands
[3] Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
[4] Department of Radiation Oncology, University Hospital Düsseldorf, Düsseldorf, Germany
[5] Medical Physics Unit, Dipartimento di Diagnostica per Immagini e Radioterapia oncologica, Fondazione Policlinico Universitario 'A. Gemelli' IRCCS, Rome, Italy
[6] Institute of Radiology, Universitá Cattolica del Sacro Cuore, Rome, Italy
[7] Radiation therapy unit, Dipartimento di Diagnostica per Immagini e Radioterapia oncologica, Fondazione Policlinico Universitario 'A. Gemelli' IRCCS, Rome, Italy
[8] Department of Radiation Oncology, University Hospital Zurich, University of Zurich, Zurich, Switzerland
[9] German Cancer Consortium (DKTK), Partner Site Munich, a Partnership between DKFZ and LMU University Hospital Munich, Germany
[10] Bavarian Cancer Research Center (BZKF), Partner Site Munich, Munich, Germany
[11] Department of Radiation Oncology, University Hospital Erlangen, Erlangen, Germany
[12] Department of Medical Physics, Ludwig-Maximilians-Universität (LMU), Munich, Germany
[*] Author to whom any correspondence should be addressed.

E-mail: **guillaume.landry@med.uni-muenchen.de**

Keywords: deep learning, MRI-linac, MRgRT, foundation models, GTV segmentation, promptable models

Supplementary material for this article is available online

Abstract

Objective. Magnetic resonance imaging (MRI) guided radiotherapy requires the delineation of gross tumor volumes (GTV) in daily MRI from MRI-linacs. Specialized models have been developed for this task for certain tumors. This study investigated an alternative, using promptable foundation models. *Approach*. Promptable foundation models were prompted with six different sparse geometric prompt types (points, boxes, 2D masks) to produce GTV segmentation masks, including Segment-anything 2 (SAM2), SAM2 fine-tuned for medical imaging (MedSAM2), and nnInteractive, an nnUnet-based promptable model for medical imaging. A diverse multi-institutional dataset of clinical GTV masks from the abdomen, lung, liver, pancreas, and pelvis sites on MRI scans from MRI-linacs was used to evaluate model outputs using various metrics, including the Dice similarity coefficient (DSC). *Main results*. The models produced segmentation masks comparable or superior to those from domain-specific models with median DSCs of up to 0.85 (nnInteractive-mask3 prompt). Prompts with more spatial information yielded better results with lower variance, with the effect reduced for nnInteractive and MedSAM2. These produced overall better results (median DSC over all prompt types 0.75 for nnInteractive, 0.70 for MedSAM2, 0.54 for SAM2). *Significance*. This investigation showed that promptable foundation models can in principle be used for GTV segmentation in MRI across multiple tumor types, although more research is necessary to reduce the variance and improve model performance.

1. Introduction

Magnetic resonance (MR) image guided radiotherapy (MRgRT) combines MR imaging (MRI) with radiation therapy linear accelerators (MRI-linac) to benefit from the soft-tissue contrast available in MRI. This improved visualization enables more precise radiation delivery and facilitates online adaptive radiotherapy, in which a treatment plan is adapted to the patient in an online or even real-time adaptive manner (Kurz *et al* 2020, Keall *et al* 2022).

A critical, time-consuming, and potentially biased task in such workflows is the manual segmentation of gross tumor volumes (GTVs), and neighboring organs-at-risk (OAR) for treatment planning. An accurate and reproducible delineation of these volumes is essential to ensure adequate tumor coverage while minimizing the irradiation of surrounding healthy tissues. For these reasons, contouring and treatment plan adaption still accounted for 27% of the total treatment time in an optimized parallel clinical workflow (Votta *et al* 2024).

To reduce manual delineation time and accelerate treatment workflows, automatic segmentation methods for OARs have been increasingly and successfully proposed and implemented in MRgRT (Eidex *et al* 2023, Psoroulas *et al* 2025). These are typically implemented for specific anatomical sites such as the pelvis (Ding *et al* 2022, Kawula *et al* 2022, Nachbar *et al* 2024, Vagni *et al* 2024, De Benetti *et al* 2025, Delopoulos *et al* 2025), the abdomen (Fu *et al* 2018, Kawula *et al* 2024, Zhou *et al* 2024) or the lung (Chekroun *et al* 2023, Ribeiro *et al* 2023).

Contrary to OARs, efforts to similarly automate the segmentation of GTVs at MRI-linacs have generally been less successful, with worse metrics compared to the segmentation of OARs (Eidex *et al* 2023). Especially since GTVs exhibit far greater heterogeneity in size, shape, location, and appearance compared to OARs, their segmentation is not only more difficult, time-consuming and resource-intensive for human experts, but it is also plagued by substantial variability between and within observers (Mercieca *et al* 2020). This variability arises from the inherent difficulty in interpreting tumor boundaries, which are often indistinct due to low image contrast, complex tumor morphology, and infiltration into adjacent tissues. For this reason, the accuracy and consistency of the GTV delineation have been described as a bottleneck in the high-precision radiation therapy chain (Mercieca *et al* 2020).

Still, automatic segmentation methods for GTVs have been proposed for MRgRT, again typically using domain/site-specific models or models specific for individual patients (Breto *et al* 2022, Kawula *et al* 2022, Li *et al* 2022, Wei *et al* 2025b).

Recently, the emergence of promptable foundation models has introduced a new paradigm in artificial intelligence (Bommasani *et al* 2022). These large-scale models, pre-trained on vast, diverse datasets, acquire a broad understanding that can be adapted to specific downstream tasks with minimal or no task-specific training, a capability known as zero-shot or few-shot learning. In computer vision, the Segment Anything Model (SAM) exemplified this by introducing promptable segmentation, where a user can guide the model to segment any object using simple geometric prompts like points, bounding boxes, or masks (Kirillov *et al* 2023).

However, the performance of foundation models trained on natural/photographic images does not necessarily transfer to medical images such as those produced by MRI scanners, due to a significant domain change. This has spurred the development of promptable foundation models adapted specifically for the biomedical domain, either by fine-tuning generalist models on medical imaging data (Lee *et al* 2024, Ma *et al* 2024a) or by training new architectures from scratch on large medical imaging datasets (Lee *et al* 2024, Isensee *et al* 2025). These specialized models promise to combine the flexibility of prompting with the domain-specific knowledge required for challenging medical tasks.

This study investigates the viability of using such promptable foundation models for GTV segmentation in MR images from MRI-linacs. A comprehensive comparison of different foundation models was conducted using various types of sparse geometric prompts. The evaluated foundation models ranged from general-purpose to those specialized for medical imaging. The evaluated prompt types range from low information single points to high information precise 2D contours in three orthogonal planes across the GTV. By evaluating these approaches on a large, multi-institutional, and multi-anatomical dataset, their performance for the task of GTV segmentation was quantified and failure modes could be identified. This allowed identifying promising prompt/model combinations, and for the assessment of their potential to become a practical tool for streamlining the MRgRT workflow and guiding future development by identifying prompt types that can be used successfully, even with the limitations of currently available foundation models.

2. Methods

2.1. Prompting in 3D

In this study, user prompts were simulated by sparsely sampling the ground truth GTV masks. To ensure comparability and compatibility between different promptable foundation models, six distinct generalized sparse representations were defined as base prompts. Each represents a plausible strategy for providing sparse segmentation guidance within a 3D volume, as follows:

- **Mask3**: Three orthogonal binary segmentation masks of the target, placed in the axial, sagittal, and coronal planes that intersect the center of the GTV.
- **Box3**: A single 3D bounding box that covers the entire GTV.
- **Points7**: One positive point at the center of the GTV, with six additional points positioned along the positive and negative directions of the three Cartesian axes, outside of the target volume at a distance of 10 pixels/voxels to the respective border of the bounding box covering the target.
- **Mask1**: A single binary segmentation mask of the GTV in the axial plane through its center.
- **Box1**: A single 2D bounding box around the GTV in the axial slice through its center.
- **Point**: A single positive point in the geometric center of the GTV.

These base prompts were applied in various configurations with the evaluated foundation models, as illustrated in figure 1. Each type of prompt conveys a different amount of spatial information about the GTV, with the information content $I(p)$ of a prompt p generally following the hierarchy:

$$\begin{array}{ccccc} I(\text{Mask3}) & \gtrsim & I(\text{Box3}) & \gtrsim & I(\text{Points7}) \\ > & & > & & > \\ I(\text{Mask1}) & \gtrsim & I(\text{Box1}) & \gtrsim & I(\text{Point}). \end{array}$$

In this study, four publicly available, promptable foundation models were investigated, representing a diverse set of model architectures, strategies to implement prompting, and data used to train the foundation models. For each model and generalized prompt type, a combination strategy was developed. When multiple plausible implementations of a given prompt type existed, a validation set was used to select the approach that produced the best performance. The same was done for hyperparameters. Options for combining outputs from multiple models or multiple prompt types iteratively were not considered due to the exponential increase in possible combinations. The foundation models used were not fine-tuned, and only publicly available checkpoints were used.

The prompting strategies for each of the four models are visually summarized in figure 1 and are explained in the following subsections. In total, 19 unique combinations of prompt types and models were considered, which are listed in table 1, with their respective identifiers following the scheme *'foundation model'-'prompt type'*.

Segment Anything Model 2 (SAM2). (SAM2) (Ravi *et al* 2024) is a general-purpose vision foundation model designed for promptable image and video segmentation tasks. It employs a modular transformer-based architecture composed of four components: (1) an image encoder, (2) a prompt encoder, (3) a mask decoder, and (4) a memory encoder and memory-attention mechanism. The image encoder encodes images to an embedding. The prompt-encoder encodes geometric prompts such as (positive and negative) points, 2D bounding boxes, and (more or less precise) masks to embeddings as well. Both are concatenated and passed as input to the mask decoder, which produces segmentation logits that are converted to binary segmentation masks by applying a threshold. For this purpose a constant threshold can be used, or, as was done for this study, a dynamic threshold, optimized to maximize the segmentation accuracy on the prompted slice following (Blöcker *et al* 2024). Finally, the memory-encoder encodes the output of the mask-decoder, and the embedding is stored in the memory-bank. For propagation to subsequent neighboring slices, the embeddings from the image encoder are passed through the memory-attention mechanism, applying attention based on the embeddings from the memory bank. This allows for mask propagation through multiple images, i.e. videos.

Although originally developed and trained for 2D photographic images and videos, SAM2 has been adapted and fine-tuned to segment structures in 3D volumetric medical images (Zhu *et al* 2024, Ma *et al* 2024b). This is typically achieved by prompting in a central slice, propagating segmentation masks bidirectionally to adjacent slices, and subsequently combining this segmentation to produce coherent 3D volumetric masks. For the initialization, 3 different types of prompts are supported: (positive and negative) point prompts, bounding box prompts, and mask prompts. To make use of the prompts with orthogonal information, the prompts are used to segment the target in each orthogonal plane and propagate bidirectionally to parallel planes on the same axes.

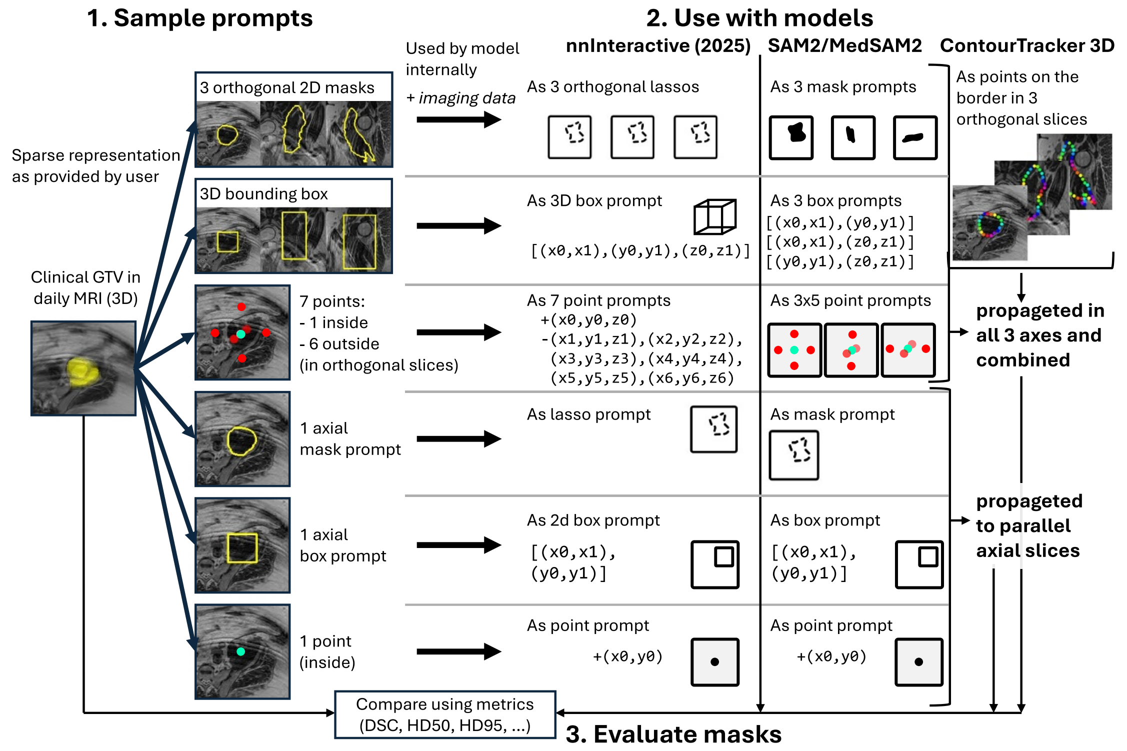

Figure 1. Overview of the prompt generation and study design. The ground-truth, clinical GTV label is used to extract six sparse prompt types, namely orthogonal masks (Mask3), 3D bounding boxes (Bo × 3), point sets (Points7), and single-point (Point) or axial mask (Mask1) or box prompts (Box1). These are converted into model-specific formats for the evaluated foundation models. The output of these models when prompted with the imaging data and these prompts are evaluated against the ground truth using metrics including DSC, HD50, and HD95.

Table 1. Overview of evaluated models with the respective identifiers (ID), the underlying base foundation model, the generalized prompt type it processed and a short description of how prompts of this type are used to instruct the underlying base model.

Base model	Base prompt	Model ID	Prompt usage
SAM2	Mask3	SAM2-mask3	as 3 mask prompts individually
	Box3	SAM2-box3	as 3 box prompts individually
	Points7	SAM2-points7	as 3 × 5 point prompts individually
	Mask1	SAM2-mask	as mask prompt
	Box1	SAM2-box	as box prompt
	Point	SAM2-point	as point prompt
MedSAM2	Mask3	MedSAM2-mask3	as 3 mask prompts individually
	Box3	MedSAM2-box3	as 3 box prompts individually
	Points7	MedSAM2-points7	as 3 × 5 point prompts individually
	Mask1	MedSAM2-mask	as mask prompt
	Box1	MedSAM2-box	as box prompt
	Point	MedSAM2-point	as point prompt
nnInteractive	Mask3	nnI-mask3	as 3 lasso prompts (combined)
	Box3	nnI-box3	as 3D box prompt (combined)
	Points7	nnI-points7	as 7 points
	Mask1	nnI-mask	as lasso prompt in axial plane
	Box1	nnI-box	as box prompt in axial plane
	Point	nnI-point	as point prompt
Contour-Tracker3D	Mask3	ConT-mask3	as points outlining the 3 masks in orthogonal slices individually

The resulting three 3D segmentation masks $M_i \in \{0,1\}^{H\times W\times Z}$ with $i \in \{\text{Sagittal}, \text{Coronal}, \text{Axial}\}$ are then combined by computing the intersection

$$M = \bigcap_i M_i \tag{1}$$

of the masks M_i.

For this study, the `SAM2.1-small` was used as a baseline instance of a general-purpose promptable segmentation model.

MedSAM2 is a version of SAM2, specifically the variant `SAM2.1-TINY`, fine-tuned for use with medical imaging, specifically 3D CT, PET and MRI images, as well as ultrasound and endoscopy imaging (Ma *et al* 2025). To segment 3D structures, an initial segmentation is performed on a single slice (usually in the axial plane) and the resulting masks are then propagated bidirectionally as described above for SAM.

This fine-tuning led to enhanced segmentation accuracy in medical applications compared to the baseline SAM2 model. Furthermore, it also improved the performance of the best SAM-based model from a 2024 CVPR competition comparing alternative SAM-based models for 3D medical imaging segmentation, EfficientMedSAMs (Ma *et al* 2024c).

In this study, one of the most recent available MedSAM2 checkpoints, `MedSAM2_MRI_LiverLesion` from April 2025, was used as a representative of fine-tuned SAM2-based models for medical imaging.

nnInteractive is another promptable foundation model, built specifically for 3D volumetric medical imaging and without making use of a transformer architecture. Instead, nnInteractive is built on nnUnet incorporating an interactive prompting mechanism using additional channels. These channels encode both positive and negative prompts across the 3 supported types (point, box/lasso, scribble) as well as the current predicted segmentation (Isensee *et al* 2025). nnInteractive was trained from scratch on a large dataset of (manually) labeled medical images, with additional pseudo-labels generated using the SAM and SAM2 models for unsupervised label generation, in a component called *SuperVoxel*. Notably, unlike SAM-based models, nnInteractive integrates prompts early in the network architecture, via convolutional operations in the first layer. In contrast, SAM2-based models combine prompts and images in the final mask-decoder component, after separately encoding both inputs into embeddings.

nnInteractive was included as representative of a recent novel approach to promptable foundation models, not based on a SAM-inspired transformer architecture, and was trained from scratch for 3D medical imaging. Thus, it is considered the state-of-the-art promptable foundation model for the task of promptable segmentation in 3D medical imaging.

nnInteractive supports six 3D input channels for the supported prompt types. These are point, bounding box/contour, and scribble prompts, in positive and negative variants, respectively. In addition, the model has two more input channels: One for the 3D image and one that may contain a current mask prediction to refine with the provided prompts, meaning that the model can be used to iteratively refine its prediction. The model outputs a binary segmentation mask. By providing this mask as input to the current mask prediction channel, iterative improvement is achieved.

While the point prompts are used similarly as with SAM-type models, the bounding and scribble prompts are implemented differently. For the bounding prompt input channel, the model was provided with 2D bounding boxes or contours (so-called lassos) containing the target to segment in a given plane during training. The model was thus pre-trained to prefer to stay within the provided contour. For the scribble prompt channel, the model was provided with a scribble inside the target, and was thus pre-trained to produce a segmentation mask that may also include parts outside of the provided scribble (Isensee *et al* 2025). In this study, the scribble channel was not used. Instead, the bounding channel was used for both mask and box prompts, with the mask prompt being dilated by 1 pixel, and their contours used as lassos prompts.

For Mask3 and Box3 prompts, orthogonal 2D prompts were provided to the model sequentially. Starting with an empty current mask prediction channel, the model is prompted with the axial prompt. With the resulting mask in the current mask prediction channel, the model is further prompted with the sagittal prompt and, finally, the coronal prompt, to produce the final output. Multiple point prompts (Points7) were also provided to the model sequentially.

Contour point tracking in 3D is a novel 3D variant of the ContourTracker model introduced by Blöcker *et al* (2024). This model is not a foundation model for image or video segmentation; instead, it uses CoTracker (Karaev *et al* 2023), a promptable foundation model for the track-any-point (TAP) task (Doersch *et al* 2023). By propagating points along a provided contour from one frame or 2D slice to another and treating these points as the vertices of a polygon, a mask propagation mechanism analogous to SAM2 is achieved.

Although this approach demonstrated slightly inferior performance compared to the large SAM2 model for MRgRT target tracking in 2D+t cine MRI data (Blöcker *et al* 2024), it offers an interesting alternative to established segmentation-based approaches. Although the contour tracking approach features significantly reduced prompting capabilities, requiring the provision of points on the contour and thus supporting only mask prompts, its inclusion complements the model selection by internally using a completely different type of promptable foundation model.

For this purpose, the approach for contour point tracking was extended to volumetric data. This was achieved in a similar fashion to the orthogonal Mask3 prompt with SAM2 and MedSAM2. For a center slice through the GTV in each plane, the segmentation mask is converted to a list of points. These points are then propagated bidirectionally to parallel slices of the same orientation, using the underlying CoTracker3 point tracking model. This is done for all planes, and the intersection of the resulting masks (equation (1)) is computed.

2.2. Dataset

For the evaluation of the promptable foundation models, a large dataset of planning MRIs from MRI-linacs was assembled, featuring tumors from patients treated at different institutions. No new data was generated for this study. Instead, existing retrospectively collected MRI scans and clinically used GTV segmentation masks were used for the evaluation.

The bulk of the dataset was retrospectively collected from patients treated with a 0.35 T MRI-linac at the Department of Radiation Oncology of the LMU Munich University Hospital (Institute I-1) with the approval of the ethics board (study number 23-1043). Informed consent was obtained. The research was carried out according to the principles embodied in the Declaration of Helsinki and according to local statutory requirements. The GTVs from the liver and abdomen sites originated exclusively from Institute I-1.

GTV segmentations of lung tumors from one other institution acquired using two 0.35 T MRI-linac systems were included in the dataset. The Amsterdam University Medical Centers (Institute I-2) provided 18 cases with the approval of the medical ethics committee (reference number 2018.602). Informed consent was obtained. The research was carried out according to the principles embodied in the Declaration of Helsinki and according to local statutory requirements.

For pancreatic tumors, the dataset used was extended to include the public dataset for Task 2 of the *PANTHER* challenge (Betancourt Tarifa *et al* 2025). This dataset consists of 50 GTVs from patients with confirmed primary pancreatic tumors who received radiotherapy using a 1.5 T Elekta Unity MRI-linac system at the Odense University Hospital (Institute I-3) in Denmark.

Finally, GTV segmentations of pelvic tumors acquired using 0.35 T MRI-linac systems were included. These were collected at the following institutions, with the approval of the respective ethics boards: LMU University Hospital (Institute I-1; study number 21-0662), Fondazione Policlinico Universitario *Agostino Gemelli* (Institute I-4; authorization number 3460), University Hospital Zurich (Institute I-5; study number 2021-D0032), and Heidelberg University Hospital (Institute I-6). Part of the pelvis cases from Institute I-1 and all pelvis cases from Institutes I-5, and I-6 were originally collected in the context of the SMILE trial with patients initially treated for prostate tumors (Fink *et al* 2024). Informed consent was obtained. The research was carried out according to the principles embodied in the Declaration of Helsinki and according to local statutory requirements.

Overall, the dataset comprised 631 annotated GTVs from 552 patients, spanning five different anatomical sites. In cases where the GTV segmentation consisted of multiple non-connected components, the individual components were treated as separate ground truth targets. This affected 52 cases of which 35 cases contained two, 9 contained three, 6 contained four, and 2 contained five distinct, non-connected regions in the clinical binary segmentation mask. Metastases were treated like primary GTVs. Segmentation masks of irradiated lymph nodes were not included in the dataset.

MRI images were acquired using various imaging parameters and acquisition settings, reflecting differences in equipment and protocols across involved institutions. The dataset featured various spatial resolutions, including: $0.89 \times 0.89 \times 2.0\,\mathrm{mm}^3$, $1.5 \times 1.5 \times 1.5\,\mathrm{mm}^3$ and $1.63 \times 1.63 \times 3.0\,\mathrm{mm}^3$. The median and most common spatial resolution was $1.5 \times 1.5 \times 1.5\,\mathrm{mm}^3$. The images from 0.35 T MRI-linacs were acquired using balanced steady-state free precession sequences (bssfp). The images from the 1.5 MRI-linac (I-3) were T2-weighted images. No universal resampling or other preprocessing steps were performed, but the models performed internal normalization, resizing, or rescaling, as explained in their perspective publications. The size distribution of the delineated GTV volumes from the same anatomical area from different institutions is shown in the supplementary materials (section 1). Only the pelvis GTVs from I-4 were significantly smaller compared to those from the other three institutions.

Table 2. Summary of the compiled GTV data. The table lists the anatomical site, origin institution, MRI-linac system (with 0.35 T referring to the ViewRay MRIdian system and 1.5 T to the Elekta Unity system) and MR acquisition parameters (balanced steady-state free precession sequences (bssfp) or T2-weighted (T2)) along with the number of patients and corresponding gross tumor volumes (GTVs) included per site.

Site	Origin	MRI-linac	N_{Patients}	N_{GTV}
Abdomen	Institute I-1	0.35 T (bssfp)	38	56
Liver	Institute I-1	0.35 T (bssfp)	91	110
Lung	Institute I-2	0.35 T (bssfp)	18	18
	Institute I-1	0.35 T (bssfp)	165	197
	Σ		183	215
Pancreas	Institute I-1	0.35 T (bssfp)	45	53
	Institute I-3	1.5 T (T2)	50	51
	Σ		95	104
Pelvis	Institute I-1	0.35 T (bssfp)	40	40
	Institute I-4	0.35 T (bssfp)	73	74
	Institute I-5	0.35 T (bssfp)	22	22
	Institute I-6	0.35 T (bssfp)	10	10
	Σ		145	146
Σ			552	631

From the cases originating at Institute I-1, a validation set of 15 GTVs from 15 different patients was randomly sampled, including 3 abdomen, 4 liver, 5 lung, and 3 pancreas GTVs. This validation set was used during development and for the initial selection of the models. The remaining GTV segmentations were used as the test set for the evaluation. The number of GTVs for each origin and anatomical site in the full dataset are listed in table 2.

2.3. Comparative study

A comparative study was conducted, comparing all the investigated models and prompt types shown in figure 1 and listed in table 1 in their ability to reproduce the ground truth GTV segmentation from their respective base prompts. For each case in the test dataset, each of the six sparse geometric prompt types proposed was sampled from the ground truth GTV segmentation mask and passed along with the imaging data to each of the 19 model-prompt combinations.

To evaluate the accuracy and performance of the model for predicting GTV segmentation masks, several metrics were computed by comparing the model outputs with the ground truth clinical GTV segmentations. The metrics were included as geometric metrics: Dice similarity coefficient (DSC) (Zou *et al* 2004), the surface DSC (3 mm margin), and the Hausdorff distance 95 (HD95), defined as the 95th percentile of the distances between the contour vertices (Huttenlocher *et al* 1993). In addition, recall and precision were computed for each case to indicate coverage of the GTV and non-coverage of the non-GTV regions, respectively. The implementation of these metrics used the monai package (MONAI Consortium 2025).

To aggregate the metrics for a given model over a group of cases, the median and interquartile ranges (IQRs) of a given metric were calculated, since the metrics did not show normality and thus non-parametric statistical tests were used.

To compute the statistical significance of the differences between the results groups for multiple cases, i.e. multiple model and prompt type combinations, the DSC and HD95 metrics were combined into a single vector and a nonparametric Friedman test (Friedman 1937) with a significance level of $\alpha = 0.05$ was applied. Where necessary, a post hoc Nemenyi test (Nemenyi 1963) was conducted to identify statistical differences between models in pairwise comparisons.

Based on these results, promising model-prompt combinations with a median DSC $\geqslant 0.8$ and a median HD95 $\leqslant 5.0$ mm were identified and their behavior was further analyzed. These thresholds were chosen as a surrogate measure of practical usefulness, to distinguish comparatively well-performing models-prompt-combinations form the rest.

The quality of their predictions was analyzed in relation to the anatomical site of the tumor and the origin institution. Here, a nonparametric Kruskal–Wallis test (Kruskal and Wallis 1952) with a significance level of $\alpha = 0.05$ was applied analogously to the Friedman test above. A Bonferroni correction $\alpha' = \alpha/k$ was applied to account for the number of tests k where applicable, to rule out incidental findings. Where necessary, a post hoc Dunn test (Dunn 1964) was performed to identify differences between the results of the different groups in pairwise comparisons.

The site-specific performance of the model-prompt combinations was compared with the reported performance of fully automatic site-specific convolutional neural networks(CNNs) and inter-observer variability from literature.

All calculations were performed on the same hardware using an RTX A6000 graphics processing unit (NVIDIA, USA). The runtime of each model was measured. However, it should be noted that performance optimization such as pytorch model compilation or use of lower-precision floating point numbers were not implemented.

2.4. Evaluation of the impact of prompt variation

To assess the robustness of the models to variations and spatial inaccuracies in geometric prompts, an additional experiment was conducted that explicitly evaluates the effect of imperfect prompts. The goal of this analysis was to determine how deviations from ideal, ground-truth prompts influence segmentation performance, thereby evaluating the sensitivity of each model-prompt combination to realistic inaccuracies in prompt definition.

These experiments were performed on the validation dataset. Ground-truth labels used to derive prompts were perturbed using randomly sampled elastic deformations and global shifts to simulate spatial inaccuracies in prompting, similarly to those human users may produce. For this purpose, the `Rand2DElastic` function from monai (MONAI Consortium 2025) was used to vary the ground-truth prompts with displacement magnitudes of up to 5 pixels, followed by an additional random global shift of up to 3 pixels.

We calculated the DSC between the output obtained when prompting the model with a given perturbed prompt and the output obtained using the non-perturbed version of the prompt. This enabled an assessment of performance degradation as a function of prompt deviation, which we estimating using the intra-prompt DSC (perturbed vs non-perturbed 2D prompts). For orthogonal prompts such as Mask3 prompts, the intra-prompt DSC was computed as average intra-prompt DSC across the three orthogonal planes.

3. Results

3.1. Quantitative overall results

Table 3 shows the metrics resulting from the comparative study. In general, the presented methods based on promptable foundation models produced segmentation masks within a wide range of qualities, from nnI-mask3 (median and IQR of the DSC over all test cases $0.85\ (0.80-0.89)$) down to sam2-point (DSC $0.13\ (0.01-0.55)$). The comparison of metrics between prompt types showed that higher information prompts produced better results, with all aggregated results per model following the expected trends: Mask3 > Box3 > Points7, Mask > Box > Point, as well as Mask3 > Mask, Box3 > Box, and Points7 > Point. In several cases, some models produced empty segmentation masks, particularly when the corresponding GTVs were small or had complex geometry. This issue was observed in ConT-mask3 (2 cases), medsam2-box3 (2 cases), medsam2-mask (5 cases), sam2-box3 (3 cases), and sam2-mask3 (3 cases). In these instances, the logits produced by the (Med)SAM2 foundation model did at no point in the image exceed the threshold applied to obtain a binary segmentation masks and/or the intersection of the masks obtained from the prompts in orthogonal slices was empty.

As is particularly evident in the HD95 metric, none of the models evaluated produced high-quality masks in all cases, and they should not be used in an unsupervised fashion. Furthermore, only the nnI-mask3, MedSAM2-mask3, and ContT-mask3 model-prompt combinations produced segmentation masks that exceeded the threshold of median DSC $\geqslant 0.8$ and HD95 $\leqslant 5.0$ mm. Thus, only these model-prompt combinations were considered further.

Comparing the runtime of the different models showed that nnI-mask3 is the fastest with a median runtime of 1.2 s, while MedSAM2-mask3 required 4.8 s and ConT-mask3 required 40.1 s. These differences originate from the way the different models operate. While nnInteractive runs inference on a volumetric image in a single pass, MedSAM2 propagates 2D contours slice by slice. Finally, ConT-mask3 propagates not full 2D contours but individual boundary points. Although not considered in this study, optimized methods could parallelize the latter two to obtain improved performance

Statistical analysis using the Friedman test revealed significant differences between the 19 model-prompt combinations. Among the highest performing combinations, the only pair without statistically significant differences was MedSAM2-mask3/ConT-mask3. Since there were no statistically significant differences between the performance of MedSAM2-mask3 and ConT-mask3, only the faster MedSAM2-mask3 was considered further along with nnI-mask3.

Table 3. Quantitative evaluation metrics for all tested model and prompt type combinations listed in table 1 aggregated over all cases in the test set. The table reports the median Dice similarity coefficient (DSC), surface DSC at a 3 mm tolerance, 95th percentile Hausdorff distance (HD95) in millimeters, as well as recall and precision. The inter quartile range is shown in braces. Results are sorted in descending order of DSC to highlight the highest-performing methods. Higher values indicate better performance for DSC, surface DSC, recall, and precision, while lower values are preferable for HD95. Finally, the median runtime over all cases from the validation set is reported. A horizontal line indicates the performance threshold used to select the highest performing combinations.

Model	DSC	Surface DSC (3 mm)	HD95 [mm]	Recall	Precision	Runtime [s]
nnI-mask3	0.85 (0.80–0.89)	0.94 (0.87–0.98)	3.4 (3.0–6.0)	0.86 (0.81–0.90)	0.87 (0.78–0.92)	1.2
MedSAM2-mask3	0.82 (0.75–0.86)	0.88 (0.78–0.96)	4.7 (3.3–7.7)	0.82 (0.70–0.90)	0.86 (0.79–0.93)	4.8
ConT-mask3	0.81 (0.74–0.86)	0.91 (0.83–0.95)	4.2 (3.4–6.9)	0.74 (0.62–0.81)	0.92 (0.90–0.94)	40.1
nnI-box3	0.79 (0.69–0.85)	0.85 (0.72–0.94)	4.7 (3.4–10.1)	0.82 (0.70–0.90)	0.83 (0.71–0.90)	0.9
nnI-mask	0.76 (0.68–0.83)	0.82 (0.67–0.93)	6.0 (3.4–12.67)	0.72 (0.60–0.81)	0.91 (0.80–0.96)	0.7
sam2-mask3	0.76 (0.62–0.85)	0.83 (0.69–0.92)	6.0 (3.7–10.8)	0.66 (0.47–0.78)	0.94 (0.90–0.97)	26.7
MedSAM2-box3	0.75 (0.66–0.82)	0.79 (0.65–0.90)	6.0 (3.6–10.5)	0.74 (0.59–0.85)	0.81 (0.73–0.89)	4.2
nnI-box	0.73 (0.61–0.82)	0.80 (0.57–0.92)	6.0 (3.4–12.7)	0.85 (0.70–0.93)	0.77 (0.54–0.91)	0.6
sam2-box3	0.71 (0.56–0.81)	0.77 (0.61–0.89)	6.5 (4.5–11.6)	0.60 (0.42–0.73)	0.92 (0.84–0.97)	26.9
MedSAM2-mask	0.69 (0.55–0.79)	0.72 (0.53–0.89)	9.0 (6.0–21.0)	0.79 (0.60–0.90)	0.77 (0.55–0.86)	2.5
MedSAM2-box	0.67 (0.53–0.78)	0.69 (0.49–0.88)	9.1 (6.0–21.2)	0.77 (0.60–0.89)	0.74 (0.49–0.88)	2.2
nnI-points7	0.66 (0.51–0.76)	0.67 (0.42–0.86)	8.6 (4.5–18.0)	0.60 (0.44–0.73)	0.94 (0.74–0.99)	2.04
MedSAM2-points7	0.65 (0.50–0.74)	0.65 (0.44–0.83)	8.6 (5.17–15.0)	0.57 (0.41–0.71)	0.91 (0.74–0.98)	3.8
nnI-point	0.65 (0.43–0.76)	0.64 (0.34–0.85)	10.6 (4.6–21.5)	0.69 (0.53–0.87)	0.89 (0.48–0.98)	0.6
sam2-points7	0.61 (0.38–0.73)	0.58 (0.37–0.79)	12.0 (6.0–22.1)	0.59 (0.42–0.73)	0.87 (0.52–0.97)	30.4
MedSAM2-point	0.58 (0.38–0.71)	0.57 (0.35–0.78)	12.4 (6.0–27.1)	0.67 (0.50–0.85)	0.69 (0.35–0.90)	2.3
sam2-mask	0.42 (0.18–0.68)	0.38 (0.17–0.72)	44.0 (12.0–113.5)	0.88 (0.74–0.95)	0.29 (0.10–0.68)	4.9
sam2-box	0.40 (0.16–0.66)	0.36 (0.16–0.65)	51.5 (12.1–117.5)	0.82 (0.68–0.91)	0.28 (0.09–0.68)	7.1
sam2-point	0.13 (0.01–0.55)	0.12 (0.00–0.45)	118.3 (31.4–201.3)	0.90 (0.72–1.00)	0.07 (0.00–0.51)	11.6

3.2. Evaluation of the impact of prompt variation

The analysis of prompt robustness (figure 2) demonstrates that the nnI-mask3 approach maintains segmentation performance comparable to that obtained with unperturbed prompts as long as the deviations of the prompt remain moderate. Meaningful decreases in accuracy were observed only when the intra-prompt DSC fell to low values.

Interestingly, for some perturbed prompts, the resulting segmentation masks showed a higher DSC compared to that obtained from the ground-truth prompts, suggesting that these prompts were not always optimal for guiding the model.

Equivalent results for the medsam2-mask3 model are provided in the supplementary materials (section 2).

3.3. Qualitative results

The most common failure modes of GTV segmentation were identified to be models following other visible structures instead of the GTV, where the other visible structure had a higher contrast to surrounding tissue compared to the GTV. Details and examples can be found in section 3 of the supplementary materials.

Despite these shortcomings, nnI-mask3 and medsam2-mask3 generally produced good overall results. Examples are shown in figure 3.

3.4. Segmentation per anatomical site

To evaluate cross-site generalization, per case metrics for nnI-mask3 and medsam2-mask3 were aggregated by anatomical site and are reported in figure 4.

The nnI-mask3 model-prompt combination produced the best results for all anatomical sites, achieving median and IQR DSCs of $0.83\ (0.76 - 0.87)$ for abdomen, $0.86\ (0.81 - 0.89)$ for liver, $0.85\ (0.80 - 0.88)$ for lung, $0.80\ (0.76 - 0.85)$ for pancreas and $0.88\ (0.83 - 0.91)$ for pelvis. The medsam2-mask3 model-prompt combination produced results of similar accuracy, with median and IQR DSCs of $0.82\ (0.73 - 0.86)$ for abdomen, $0.83\ (0.74 - 0.88)$ for liver, $0.82\ (0.73 - 0.87)$ for lung, $0.80\ (0.73 - 0.83)$ for pancreas, and $0.83\ (0.79 - 0.87)$ for pelvis.

Statistical analysis using a Kruskal–Wallis test showed that the differences between the results for the different sites were statistically significant for both. Only the lung / liver pair was not statistically significant as determined using the Dunn test.

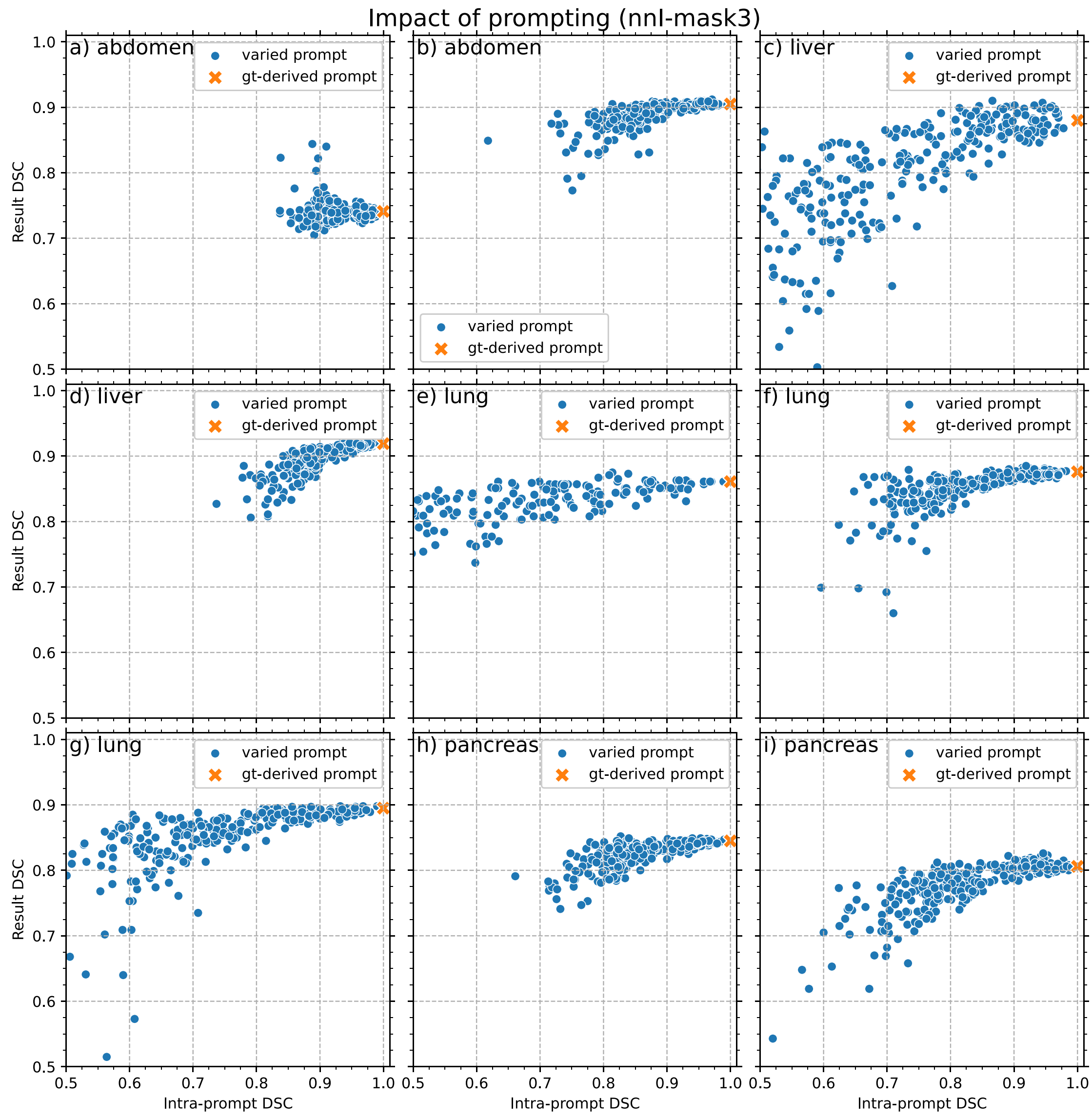

Figure 2. Impact of prompt deviations on the segmentation performance of the nnI-mask3 model-prompt combination for nine validation cases ((a)–(i)). Each point represents a single prediction either the ground-truth-derived prompt (orange) or prompts perturbed with a random elastic deformation and affine transformation (blue). The intra-prompt Dice similarity coefficient (DSC) (x-axis) quantifies the similarity between the perturbed and original 2D prompts, and the resulting DSC (y-axis) denotes the similarity between the model volumetric output and the ground-truth segmentation. The original prompt (orange) appears at $x = 1$ by definition. Across cases, segmentation accuracy remains stable for small prompt deviations but decreases as the intra-prompt DSC declines. This indicates robustness with respect to deviations from the ground-truth prompt.

3.5. Segmentation per institution

Figure 5 reports performance across GTV institution of origin to assess whether promptable foundation models can generalize across inter-institutional contouring preferences.

Statistically relevant differences between institutions were found only for the pelvic site, where the segmentation accuracy of both nnI-mask3 and medsam2-mask3 where statistically significantly higher for GTVs from Institute I-5. Notably, these inter-institutional differences did not align with the statistically significant differences of the GTV sizes, where the GTVs from Institute I-4 were different from the others. This suggests that the difference is indeed based on contouring preferences and not just size differences. The GTV segmentations from Institute I-4 and I-6 were generally confined to anatomical structures visible on MRI, such as the prostate. At Institute I-1, GTV segmentation masks often extended beyond anatomical boundaries or included spiky tumor extensions, making them more difficult to reproduce from sparse prompts. At Institute I-4, lymphatic pathways were also included for some GTVs. These do not necessarily align with structures visible on MRI and are more challenging to segment from sparse prompts. In comparison, the GTV segmentations from I-5 were more aligned with visible anatomical boundaries, resulting in the higher performance.

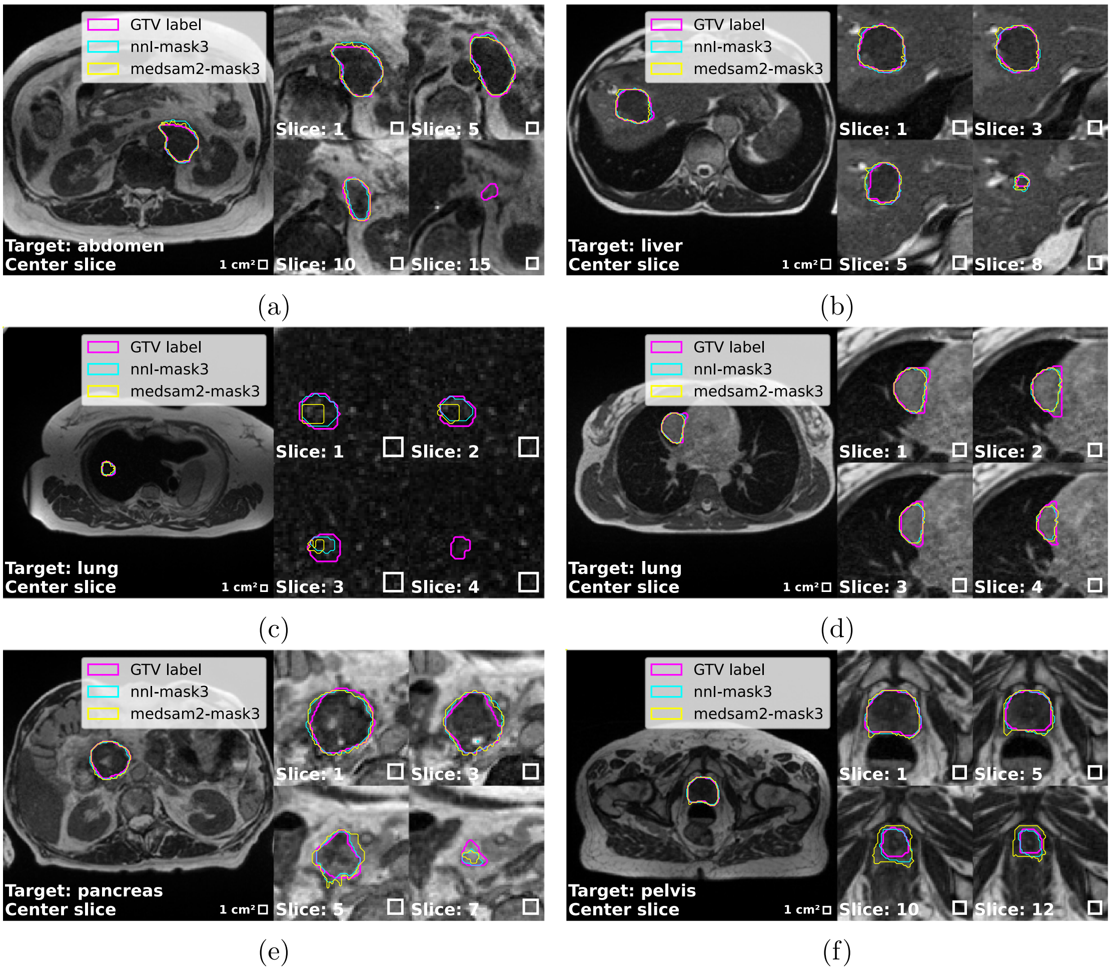

Figure 3. Selected examples of GTV delineations from the clinical ground-truth label (white) and predictions from nnI-mask3 (cyan) and medsam2-mask3 (yellow). The axial slices used for prompting (left) and selected parallel slices (right) are shown, respectively. A 1 cm^2 square in the lower right corner indicates the scaling and cropping, applied for better visibility.

No statistically significant differences were reported for the other two anatomical sites with cases from multiple institutions (lung and pancreas).

Comparing the resulting metrics for cases from the same anatomical region for different MRI resolutions did not show any statistically significant differences (Friedmann test, $p > 0.05$) for the nnI-mask3 or medsam2-mask3 model-prompt combinations (section 4 in the supplementary materials). The cases from Institute I-5 were excluded for this purpose, due to having a distinct resolution and already having been confirmed to be significantly different from the others.

4. Discussion

Overall, the results demonstrated that promptable foundation models can be successfully used to segment GTVs in MR images acquired from MRI-linacs with median DSC up to 0.85 when prompted with 2D mask prompts in three orthogonal slices.

4.1. Overall investigation

The quantitative results indicated that accurate segmentation masks require precise prompts that define the shape and boundaries of the target in orthogonal planes. Prompts that provided more spatial guidance consistently improved segmentation accuracy.

Qualitative analysis showed that GTV segmentation in MR imaging from MRI-linacs remains challenging for promptable models, primarily due to three failure modes: (1) failure to accurately detect target boundaries, (2) incorrect segmentation of unrelated anatomical structures or surrounding tissue, and (3) inability to segment microscopic components of GTVs that are difficult to distinguish on MR images. The ConT-mask3, medsam2-box3, medsam2-mask, sam2-box3, and sam2-mask3 models also produced empty segmentation masks for a small number of small or complex-shaped GTVs.

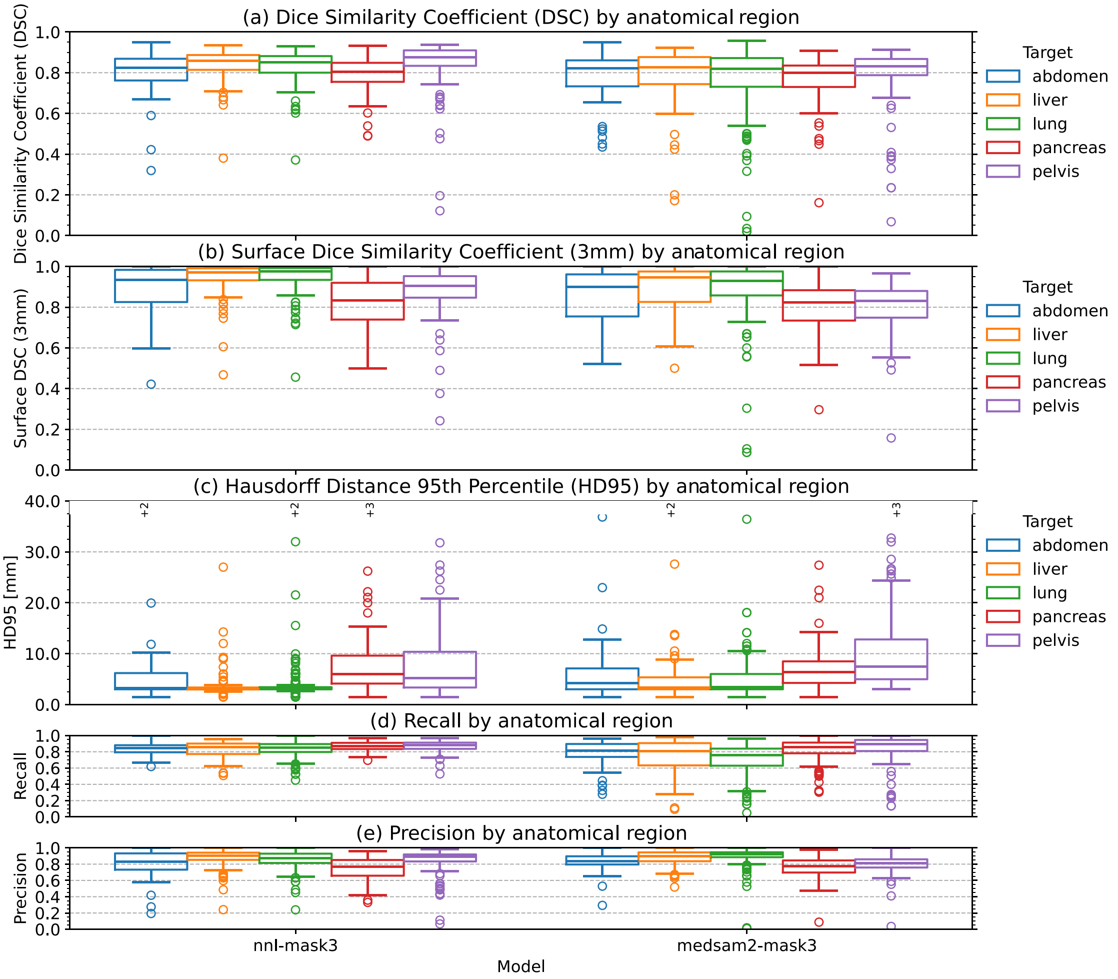

Figure 4. Visualization of the per case (a) Dice similarity coefficient (DSC), (b) surface DSC (3 mm), (c) Hausdorff distance (95th percentile), (d) recall, and (e) precision metrics by anatomical site of the target for selected model and prompt combinations. The models are sorted by prompt type and descending DSC. The box color indicates the site. The center-line of the box indicates the median, the box signifies the first and third quartiles, and the whiskers extend to the extremes of the distribution, up to 1.5 times IQR. Outliers beyond are shown as circles. The number of extreme outliers outside of the depicted axis range in (c) is stated at the top of the subplot.

The use of higher-information prompts mitigates these failure modes in various ways. The mask and bounding box prompts in orthogonal slices helped inform the foundation model about the precise extent of the GTV in each plane. Mask prompts and, to a lesser extent, bounding-box prompts reduce the occurrence of segmentation of unrelated structures or surrounding tissue. In contrast, these occur frequently with point prompts due to a potentially low contrast between the GTV and surrounding tissue relative to the contrast between surrounding tissue and other visible anatomical features. Finally, mask prompts provided detailed guidance about which regions to include, thereby improving the detection of subtle or microscopic tumor components.

Further research in the mechanisms of how prompts influence the output of machine learning models may be constructive. Similarly, research into prompt types beyond the six examined in this study may further improve the usability of promptable foundation models. Such could include scribble prompts (Wong *et al* 2024, Isensee *et al* 2025) or non-geometric prompts, such as text or natural language descriptions (Rajendran *et al* 2024, Zhao *et al* 2024).

Between the four foundation models investigated, nnInteractive consistently outperformed the others in all metrics across all prompt types.

For the type of prompt containing the highest level of information, three orthogonal masks, the approaches that produced good segmentation masks (median DSC $\geqslant$ 0.8 and HD95 $\leqslant$ 5.0 mm) were: The nnI-mask3 model (median DSC 0.85), the nnInteractive model prompted with three mask/lasso prompts, and medsam2-mask3 (DSC 0.82), merging masks of threefold prompting with masks in orthogonal planes, and the ConT-mask3 (DSC 0.81), propagating points on these masks. Due to the absence

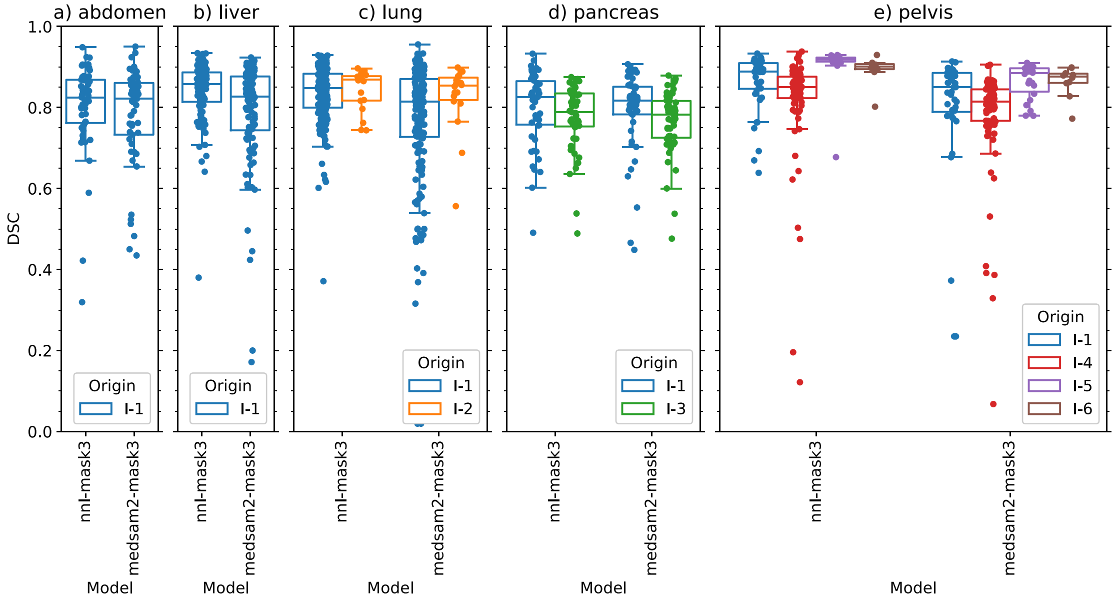

Figure 5. Visualization of the per case Dice similarity coefficient (DSC) for cases from different institutions by anatomical site. The box color indicates the origin institution. The center-line of the box indicates the median, the box signifies the first and third quartiles, and the whiskers extend to the extremes of the distribution, up to 1.5 times IQR. Individual datapoints are overlayed onto the boxes.

of a statistically significant difference between medsam2-mask3 and ConT-mask3, only medsam2-mask3 was included in further investigations, since it showed a faster runtime.

4.2. Robustness against prompt variation

The evaluation of the impact of prompt variations on the resulting segmentation accuracy showed that the nnI-mask3 and medsam2-mask3 model-prompt combinations were robust with respect to minor variations in the prompts. The presented methods thus do not require precise prompts, but ones that are reasonably close to the GTV.

Notably, for some of the varied prompts a higher segmentation accuracy was observed compared to the ground-truth prompts. The method of obtaining ground-truth prompts directly from the ground-truth label in central orthogonal slices was thus not necessarily optimal but instead reasonably good as well. Further research could investigate the actual intra-observer prompting variability.

4.3. Site specific performance

Comparison of model performance for different anatomical sites showed small performance differences for targets from different anatomical sites. These were statistically significant, with the exception of those between lung and liver (Dunn test, $p > 0.05$).

This behavior matched expectations, as tumors from different anatomical sites feature different contrast on MRI, thus producing more or less sharply defined contours. In addition, for the pelvis region, the GTV might not directly follow the prostate itself but may include additional surrounding tissue and sometimes parts of the seminal vesicles.

To put the site-specific performance of the proposed AI-assisted segmentation into context, the metrics distributions where compared to site-specific inter-observer variability and the performance of automatic segmentation models (typically CNNs from literature.

For GTVs from the pelvis region, nnI-mask3 and medsam2-mask3 achieved high segmentation accuracy with median DSCs of 0.88 and 0.83, respectively. This exceeded the inter-observer variability reported by Hearn *et al* (2021) of DSC 0.705 for prostate delineations in high-field MRI. The performance is comparable to CNN-based reference values, specifically to a baseline 3D U-net model trained specifically for pelvic segmentation (Kawula *et al* 2022), which achieved mean DSCs of 0.84 and 0.74 for two cohorts (excluding the post-prostatectomy target volumes found in that study).

For lung GTVs, nnI-mask3 and medsam2-mask3 achieved median DSCs of 0.85 and 0.83. These values are comparable to the reported inter-observer variability for lung targets on high-field MRI of 0.68 (Zhang *et al* 2022) for severe cases with atelectasis and in general up to 0.9 (Kumar *et al* 2022) against a STAPLE volume. In comparison with a CNN-based automatic approach, performance exceeds that reported by Wei *et al* (2025b) for nnUNet-based baseline models applied to 0.35 T MRI-linac fraction images with mean DSCs of 0.51 and 0.73 for two cohorts.

Segmentation of liver GTVs achieved a median DSC of 0.86 for nnI-mask3 and 0.83 for medsam2-mask3. These results are consistent with the reported inter-observer DSC of 0.82 (Marshall *et al* 2023) or 0.87 (Peltenburg *et al* 2024) for high-field MRI. A study by Li *et al* (2022) reported a mean DSC of 0.67 for a nnUNet model trained on 1.5 T MRI-linac data.

For pancreas GTVs, both nnI-mask3 and medsam2-mask3 achieved median DSCs of 0.80. This performance exceeds or is comparable to the agreement between observers reported in Caravatta *et al* (2019) for two GTVs with inter-observer DSCs 0.61 and 0.53, and in Zhang *et al* (2025) with inter-observer DSC 0.80 for full-organ pancreas delineations. The leading method in the PANTHER challenge (Betancourt Tarifa *et al* 2025) (a convolutional nnUNet) achieved a mean DSC of 0.529 for automatic delineation of pancreas GTVs in 1.5 T T2-weighted images from MRI-linacs.

Overall, these comparisons show that the AI-assisted methods presented a high segmentation accuracy comparable to inter-observer variability and generally exceed the performance of fully automated CNN-based models. This is achieved not via site-specific fine-tuning, but via prompting which directs the promptable foundation models to accurately delineate the specified target.

4.4. Segmentation per institution

The evaluation showed that model performance on GTVs from different institutions was consistent within the same anatomical site. This finding suggests that the approaches tested generalize well across variations in imaging protocols or segmentation practices. This confirmed an anticipated benefit from the pretrained foundation models.

For the lung site, no statistically significant differences in segmentation performance were observed between institutions for any of the model-prompt combinations.

For the pancreas, although the data originated from two different MRI-linac systems (0.35 T and 1.5 T), which could have introduced additional variability, statistical analysis similarly revealed no significant differences for nnI-mask3 or medsam2-mask3.

Finally, for the prostate site, significant differences were detected in the results for cases originating from different institutions. In particular, in the cases of Institute I-5, a consistently better performance was achieved across the evaluated models. This points to a potential systematic influence of factors such as patient selection or contouring practices specific to that institution, which resulted in more regular, for foundation models ‘easier to reproduce’, segmentation masks. These findings underscore that, while promptable foundation models can perform at a high level across anatomical sites and institutions, additional harmonization efforts or model adaptation may be required for certain scenarios.

The limited differences in resolution present in the data set were found to have no statistically significant impact on performance (Friedmann test, $p > 0.05$).

5. Conclusions

This study demonstrates that promptable foundation models can be used for semi-automated target volume contouring. Their effectiveness for producing GTV segmentation masks based on sparse geometric prompts in MR images from MRI-linacs was evaluated for four different base foundation models and up to six distinct prompt types that could be provided by a user.

Among the four foundation models evaluated, nnInteractive consistently outperformed the others. When prompted with the highest-information prompt, three orthogonal masks intersecting at the center of the GTV, it produced results comparable to those produced by non-foundation models trained for specific tasks, despite failure to correctly segment the target in certain cases. Thus, none of the models presented in their current form should be used in a completely automatized workflow without a human-expert in the loop, scrutinizing and manually correcting segmentation masks. However, even with these limitations, promptable foundation models could be a valuable stand-alone or supplementary tool for AI-assisted segmentation of GTVs in MRgRT workflows.

Data availability statement

The MRI imaging data and segmentation masks used in this study cannot generally be made publicly available due to privacy concerns. However, part of the data used originated from the public dataset of the PANTHER challenge and can be obtained there. The new data generated in this study, the per-case metrics for each model and prompt combination, can be made available upon reasonable request. The code for the methods presented will be made available. The data that support the findings of this study are available upon reasonable request from the authors. Data and source code will be made available at the following URL: https://github.com/LMUK-RADONC-PHYS-RES/gtv-segmentation-with-foundation-models.

Supplementary figures available at https://doi.org/10.1088/1361-6560/ae2db9/data1.

Acknowledgments

Moritz Rabe and Maria Raedler are thanked for support related to the curation of the MRI data from institutions I-2 and I-4, respectively.

The authors would like to thank the BZKF Lighthouse on Image-Guidance in Local Therapies and Deutsche Forschungsgemeinschaft (DFG) project 469106425 for supporting the project.

The organizers of the *PANTHER* challenge are thanked for their efforts in the preparation of their challenge and the release of the public dataset.

ORCID iDs

Tom Julius Blöcker 0009-0002-2650-4680
Sebastian Klüter 0000-0003-3139-3444
Juliane Hörner-Rieber 0000-0003-3911-4438
Riccardo Dal Bello 0000-0002-8755-377X
Denis Dudas 0000-0003-0667-3727
Marco Riboldi 0000-0002-2431-4966
Guillaume Landry 0000-0003-1707-4068

References

Betancourt Tarifa A S, Mahmood F, Bernchou U and Koopmans P J 2025 PANTHER challenge: public training dataset *Zenodo* (available at: https://zenodo.org/doi/10.5281/zenodo.15192302)

Blöcker T *et al* 2024 MRgRT real-time target localization using foundation models for contour point tracking and promptable mask refinement *Phys. Med. Biol.* **70** 015004

Bommasani R *et al* 2022 On the opportunities and risks of foundation models (arXiv:2108.07258)

Breto A L *et al* 2022 Deep learning for per-fraction automatic segmentation of gross tumor volume (GTV) and Organs at Risk (OARs) in adaptive radiotherapy of cervical cancer *Front. Oncol.* **12** 854349

Caravatta L *et al* 2019 Magnetic resonance imaging (MRI) compared with computed tomography (CT) for interobserver agreement of gross tumor volume delineation in pancreatic cancer: a multi-institutional contouring study on behalf of the AIRO group for gastrointestinal cancers *Acta Oncol.* **58** 439–47

Chekroun M, Mourchid Y, Bessiéres I and Lalande A 2023 Deep learning based on efficientnet for multiorgan segmentation of thoracic structures on a 0.35 T MR-linac radiation therapy system *Algorithms* **16** 564

De Benetti F *et al* 2025 Enhancing patient-specific deep learning based segmentation for abdominal magnetic resonance imaging-guided radiation therapy: a framework conditioned on prior segmentation *Phys. Imaging Radiat. Oncol.* **34** 100766

Delopoulos N *et al* 2025 Implementation and clinical evaluation of an in-house thoracic auto-segmentation model for 0.35 T magnetic resonance imaging guided radiotherapy *Phys. Imaging Radiat. Oncol.* **35** 100819

Ding J, Zhang Y, Amjad A, Xu J, Thill D and Li X A 2022 Automatic contour refinement for deep learning auto-segmentation of complex organs in MRI-guided adaptive radiation therapy *Adv. Radiat. Oncol.* **7** 100968

Doersch C, Yang Y, Vecerik M, Gokay D, Gupta A, Aytar Y, Carreira J and Zisserman A 2023 TAPIR: tracking any point with per-frame initialization and temporal refinement (arXiv:2306.08637)

Dunn O J 1964 Multiple comparisons using rank sums *Technometrics* **6** 241–52

Eidex Z, Ding Y, Wang J, Abouei E, Qiu R L J, Liu T, Wang T Y and X 2023 Deep learning in MRI-guided radiation therapy: a systematic review *J. Appl. Clin. Med. Phys.* **25** e14155

Fink C *et al* 2024 Stereotactic ultrahypofractionated MR-guided radiotherapy for localized prostate cancer—acute toxicity and patient-reported outcomes in the prospective, multicenter SMILE phase II trial *Clin. Transl. Radiat. Oncol.* **46** 100771

Friedman M 1937 The use of ranks to avoid the assumption of normality implicit in the analysis of variance *J. Am. Stat. Assoc.* **32** 675–701

Fu Y *et al* 2018 A novel MRI segmentation method using CNN-based correction network for MRI-guided adaptive radiotherapy *Med. Phys.* **45** 5129–37

Hearn N *et al* 2021 Prostate cancer GTV delineation with biparametric MRI and 68Ga-PSMA-PET: comparison of expert contours and semi-automated methods *Br. J. Radiol.* **94** 20201174

Huttenlocher D, Klanderman G and Rucklidge W 1993 Comparing images using the Hausdorff distance *IEEE Trans. Pattern Anal. Mach. Intell.* **15** 850–63

Isensee F *et al* 2025 nnInteractive: redefining 3D promptable segmentation (arXiv:2503.08373)

Karaev N, Rocco I, Graham B, Neverova N, Vedaldi A and Rupprecht C 2023 CoTracker: it is better to track together (arXiv:2307.07635)

Kawula M *et al* 2022 Patient-specific transfer learning for auto-segmentation in adaptive 0.35 T MRgRT of prostate cancer: a bi-centric evaluation *Med. Phys.* **50** 1573–85

Kawula M *et al* 2024 Personalized deep learning auto–segmentation models for adaptive fractionated magnetic resonance–guided radiation therapy of the abdomen *Med. Phys.* **52** 2295–304

Keall P J *et al* 2022 Icru report 97 mri-guided radiation therapy using mri-linear accelerators *J. ICRU* **22** 1–100

Kirillov A *et al* 2023 Segment anything (arXiv:2304.02643)

Kruskal W H and Wallis W A 1952 Use of ranks in one-criterion variance analysis *J. Am. Stat. Assoc.* **47** 583–621

Kumar S, Holloway L, Boxer M, Yap M L, Chlap P, Moses D and Vinod S 2022 Variability of gross tumour volume delineation: MRI and CT based tumour and lymph node delineation for lung radiotherapy *Radiother. Oncol.* **167** 292–9

Kurz C *et al* 2020 Medical physics challenges in clinical MR-guided radiotherapy *Radiat. Oncol.* **15** 93

Lee H H *et al* 2024 Foundation models for biomedical image segmentation: a survey (arXiv:2401.07654)

Li Z *et al* 2022 Patient-specific daily updated deep learning auto-segmentation for MRI-guided adaptive radiotherapy *Radiother. Oncol.* **177** 222–30

Ma J *et al* 2024c Efficient MedSAMs: segment anything in medical images on laptop (arXiv:2412.16085)

Ma J *et al* 2025 MedSAM2: segment anything in 3D medical images and videos (arXiv:2504.03600)

Ma J, He Y, Li F, Han L, You C and Wang B 2024a Segment anything in medical images *Nat. Commun.* **15** 654

Ma J, Kim S, Li F, Baharoon M, Asakereh R, Lyu H and Wang B 2024b Segment anything in medical images and videos: benchmark and deployment (arXiv:2408.03322)

Marshall C *et al* 2023 Interobserver variability of gross tumor volume delineation for colorectal liver metastases using computed tomography and magnetic resonance imaging *Adv. Radiat. Oncol.* **8** 101020

Mercieca S, Belderbos J S A and van Herk M 2020 Challenges in the target volume definition of lung cancer radiotherapy *Transl. Lung Cancer Res.* **10** 1983

MONAI Consortium 2025 Monai: medical open network for AI *Zenodo* https://zenodo.org/doi/10.5281/zenodo.4323058

Nachbar M *et al* 2024 Automatic AI-based contouring of prostate MRI for online adaptive radiotherapy *Z. Med. Phys.* **34** 197–207

Nemenyi P B 1963 Distribution-free multiple comparisons *PhD Thesis* Princeton University

Peltenburg J E *et al* 2024 Interobserver variation in tumor delineation of liver metastases using magnetic resonance imaging *Phys. Imaging Radiat. Oncol.* **30** 100592

Psoroulas S, Paunoiu A, Corradini S, Hörner-Rieber J and Tanadini-Lang S 2025 MR-linac: role of artificial intelligence and automation *Strahlenther. Onkol.* **201** 298–305

Rajendran P *et al* 2024 Auto-delineation of treatment target volume for radiation therapy using large language model-aided multimodal learning *Int. J. Radiat. Oncol.*Biol.*Phys.* **121** 230–40

Ravi N *et al* 2024 SAM 2: segment anything in images and videos (arXiv:2408.00714)

Ribeiro M F *et al* 2023 Deep learning based automatic segmentation of organs-at-risk for 0.35 T MRgRT of lung tumors *Radiat. Oncol.* **18** 135

Vagni M *et al* 2024 Auto-segmentation of pelvic organs at risk on 0.35T MRI using 2D and 3D generative adversarial network models *Phys. Medica* **119** 103297

Votta C *et al* 2024 Evaluation of clinical parallel workflow in online adaptive MR-guided radiotherapy: a detailed assessment of treatment session times *Tech. Innov. Patient Support Radiat. Oncol.* **29** 100239

Wei C *et al* 2025b Deep learning-based contour propagation in magnetic resonance imaging-guided radiotherapy of lung cancer patients *Phys. Med. Biol.* **70** 145018

Wong H E, Rakic M, Guttag J and Dalca A V 2024 ScribblePrompt: fast and flexible interactive segmentation for any biomedical image *European Conf. on Computer Vision (ECCV)*

Zhang H *et al* 2022 Reduction of inter-observer variability using MRI and CT fusion in delineating of primary tumor for radiotherapy in lung cancer with atelectasis *Front. Oncol.* **12** 841771

Zhang Z *et al* 2025 Large-scale multi-center CT and MRI segmentation of pancreas with deep learning *Med. Image Anal.* **99** 103382

Zhao T *et al* 2024 A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities *Nat. Methods* **22** 166–76

Zhou Y, Lalande A, Chevalier C, Baude J, Aubignac L, Boudet J B and I 2024 Deep learning application for abdominal organs segmentation on 0.35 T MR-Linac images *Front. Oncol.* **13** 1285924

Zhu J, Hamdi A, Qi Y, Jin Y and Wu J 2024 Medical SAM 2: segment medical images as video via Segment Anything Model 2 (arXiv:2408.00874)

Zou K H *et al* 2004 Statistical validation of image segmentation quality based on a spatial overlap index *Acad. Radiol.* **11** 178–89