Medical Image Analysis 112 (2026) 104134

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

MRIgRT real-time target tracking: TrackRAD2025 challenge report

Tom Julius Blöcker [1], Pia A.W. Görts [2,3], Yiling Wang [4], Elia Lombardo [1], Adrian Thummerer [1], Yu Fan [4], Yue Zhao [4], Christianna Iris Papadopoulou [1], Coen Hurkmans [2,5,6], Rob H.N. Tijssen [2], Davide Cusumano [7], Martijn PW Intven [8], Pim Borman [8], Marco Riboldi [9], Denis Dudáš [1,10], Hilary L. Byrne [11,12], Lorenzo Placidi [13], Marco Fusella [14], Michael Jameson [11], Miguel A. Palacios [15], Paul Cobussen [15], Tobias Finazzi [15], Shyama U. Tetar [15], Cornelis J.A. Haasbeek [15], Paul Keall [12], Matteo Maspero [8,16], Christopher Kurz [1], Amparo Soeli Betancourt Tarifa [17,18], Kailin He [19], Shengqian Zhu [19], Ying Song [20], Guangjun Li [20], Junjie Hu [19], Felix Knispel [21,22], Sergios Gatidis [22], Hung Chu [23], Jiapan Guo [24], Maximilian Nielsen [25], Thilo Sentker [25], Valentin Boussot [26], Cédric Hémon [26], Jing Ni [27,28], Konstantinos Georgas [29], Theodoros P. Vagenas [29], George K. Matsopoulos [29], Guillaume Landry [1,30,31,*]

[1] *Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany*
[2] *Department of Radiation Oncology, Catharina Hospital, Eindhoven, The Netherlands*
[3] *Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands*
[4] *Department of Radiation Oncology, Radiation Oncology Key Laboratory of Sichuan Province, Sichuan Clinical Research Center for Cancer, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, University of Electronic Science and Technology of China, Chengdu, China*
[5] *Department of Electrical Engineering, Technical university Eindhoven, The Netherlands*
[6] *Department of Applied Physics and Science education, Technical university Eindhoven, The Netherlands*
[7] *Medical Physics Unit, Mater Olbia Hospital, Olbia, Italy*
[8] *Department of Radiotherapy, University Medical Center Utrecht, Utrecht, The Netherlands*
[9] *Department of Medical Physics, Ludwig-Maximilians-Universität München, Garching, Germany*
[10] *Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Prague, Czechia*
[11] *GenesisCare, St Vincent's Hospital, Sydney, Australia*
[12] *Sydney School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, Australia*
[13] *Fondazione Policlinico Universitario Agostino Gemelli, IRCCS, Department of Diagnostic Imaging, Oncological Radiotherapy and Hematology, Rome, Italy*
[14] *Department of Radiation Oncology, Abano Terme Hospital, Abano Terme, Veneto, Italy*
[15] *Department of Radiation Oncology, Amsterdam UMC location Vrije Universiteit Amsterdam, Amsterdam, The Netherlands*
[16] *Computational Imaging Group for MR Diagnostics & Therapy, University Medical Center Utrecht, Utrecht, The Netherlands*
[17] *Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, The Netherlands*
[18] *Diagnostic Image Analysis Group, Department of Medical Imaging, Radboud University Medical Center, Nijmegen, The Netherlands*
[19] *Machine Intelligence Laboratory College of Computer Science, Sichuan University, Chengdu 610065, PR China*
[20] *Department of Radiation Oncology, West China Hospital, Sichuan University, Chengdu 610041, PR China*
[21] *RWTH Aachen University, Aachen, Germany*
[22] *Stanford University, Stanford, CA, USA*
[23] *Donald Smits Center for Information and Technology, University of Groningen, Groningen, NL 9700AB, The Netherlands*
[24] *Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Nijenborg 9, Groningen, 9747AG, NL, The Netherlands*
[25] *Institute of Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
[26] *Univ Rennes 1, CLCC Eugène Marquis, INSERM, LTSI – UMR 1099, F-35000 Rennes, France*
[27] *Shanghai United Imaging Healthcare Co., Ltd., Shanghai 201815, China*
[28] *School of Biomedical Engineering, ShanghaiTech University, Shanghai, China*
[29] *Biomedical Engineering Lab (BEL), School of Electrical and Computer Engineering, National Technical University of Athens, 15780, Greece*
[30] *Bavarian Cancer Research Center (BZKF), Munich, Germany*
[31] *German Cancer Consortium (DKTK), partner site Munich, a partnership between DKFZ and LMU University Hospital Munich, Germany*

* Correspondence to: Department of Radiation Oncology, Medical Center of the University of Munich, Campus Grosshadern, Marchioninistr. 15, 81377 Munich.
E-mail address: guillaume.landry@med.uni-muenchen.de (G. Landry).

https://doi.org/10.1016/j.media.2026.104134
Received 3 December 2025; Received in revised form 24 March 2026; Accepted 10 May 2026
Available online 23 May 2026

ARTICLE INFO

Dataset link: https://trackrad2025.grand-challenge.org/, https://github.com/LMUK-RADONC-PHYS-RES/trackrad2025, https://huggingface.co/datasets/LMUK-RADONC-PHYS-RES/TrackRAD2025

Keywords:
Deep learning
Respiratory motion
MRI-linac
MRI-guidance
Motion management

ABSTRACT

Magnetic resonance imaging (MRI)-guided radiotherapy (MRIgRT) integrates MRI with linear accelerators (MRI-linacs), enabling real-time motion management based on temporally resolved 2D MRI (cine-MRI). Current systems rely on template matching or deformable image registration for radiotherapy target (typically the gross tumor volume) localization, which allows beam gating. Further advances in localization could support more precise and efficient delivery methods.

https://trackrad2025.grand-challenge.org/ was organized to provide a common dataset to benchmark algorithms for MRIgRT target tracking in 2D+t cine-MRI. Participants propagated target segmentation masks from an initialization frame across subsequent frames. The dataset comprised sagittal cine-MRI scans of 585 cancer patients undergoing radiotherapy at 0.35 T and 1.5 T MRI-linacs at six different institutions, with expert-annotated targets in 108 sequences. Target sites included the thorax (179 cases), abdomen (266 cases), and pelvis (140 cases). A total of 477 unlabeled and 50 labeled cases were provided for training purposes, 58 cases were kept private for preliminary testing (8) and final evaluation (50). The algorithms submitted by participants were executed on the challenge platform and assessed using metrics in three categories: geometric accuracy, surrogate dose accuracy and execution speed. Rankings were derived via a Rank-Then-Mean scheme.

TrackRAD2025 attracted 148 registrations from 28 countries, 100 preliminary submissions and 24 final submissions from 14 teams. The top five methods achieved mean Dice similarity coefficients >0.87 and Euclidean center distances <2.1 mm, comparable to interobserver variability. Leading top five solutions featured foundation models with (4) or without (1) finetuning. Field strength had minimal effect on performance and tracking worked better for the pelvis with reduced motion amplitude compared to the thorax and abdomen cases, which achieved equivalent performance.

TrackRAD2025 established a benchmark for MRIgRT tracking on multi-institutional cine-MRI data, highlighting foundation models as promising for clinical translation.

1. Introduction

The use of magnetic resonance imaging (MRI) to visualize and characterize motion is gaining increasing importance in cancer treatment, particularly in radiotherapy. Motion management is crucial for tumors affected by respiratory motion, such as liver, pancreatic, or thoracic tumors, to ensure a high radiation dose to the tumor and preservation of neighboring organs. The recent development of MRI-guided radiotherapy (MRIgRT), based on hybrid MRI-linear accelerator (linac) systems (MRI-linacs), offers the possibility of adapting to changes in tumor position during treatment (Keall et al., 2022). In 2D+t cine-MRI scans, 2D MRI images are recorded in fast sequence to allow for real-time motion visualization. Adaptation of radiation therapy delivery requires tumor segmentation in each time-resolved frame (Green et al., 2018; Jassar et al., 2023) to allow beam adaptation, e.g. via beam gating, where delivery is started and stopped when the target enters or exits a specific region. This needs to be done in real-time with low latency at frame rates of up to 8 Hz or beyond, with high accuracy and robustness to ensure the sparing of critical organs. Currently, clinically available solutions rely on conventional deformable image registration (DIR) (Palacios et al., 2023) or template matching (Jassar et al., 2023) to propagate contours from a labeled frame and struggle with large non-rigid motion (Mazur et al., 2016; Keiper et al., 2020). This limits treatment to beam gating, where the beam is turned off for large motion. Improved tumor tracking on cine-MRI at MRI-linacs will benefit patients with various motion-affected tumor entities, resulting in more accurate dose delivery.

The fast inference of artificial intelligence (AI) methods, obtained by shifting computation time to the training phase, has shown promise for this task (Lombardo et al., 2024). However, no large public dataset was available to foster model development and enable a fair comparison. TrackRAD2025 provided the first public, multi-institutional dataset and evaluation platform for comparing the latest developments in cine-MRI-based tumor delineation methods (Wang et al., 2025).

In this challenge report, we report the setup of the challenge, participation, evaluation, and ranking of submissions based on the specified metrics. The performance of the submissions is evaluated considering factors such as magnetic field strength and anatomical region. In addition, the stability of the ranking is assessed.

2. Material and methods

2.1. Challenge setup

The TrackRAD challenge enabled teams to evaluate tracking algorithms on a multi-institutional dataset. The challenge was hosted on the Grand Challenge platform at https://trackrad2025.grand-challenge.org. The participants could team up and submit an algorithm for automatic evaluation.

The task of TrackRAD2025 was real-time tumor tracking in time-resolved sagittal 2D cine-MRI sequences. Participants had to create algorithms that generate a tumor segmentation mask for every frame in a sequence (i.e. case), starting from a given ground truth label segmentation mask in the first frame. This task is also known as target localization, which is a pre-requisite for the effective implementation of tumor tracking approaches (Lombardo et al., 2024). This setup is illustrated in Fig. 1.

To ensure the applicability of the algorithms submitted for MRIgRT, the segmentation masks produced were automatically evaluated on the Grand Challenge platform with four geometric metrics and one dose-based metric. Runtime restrictions were implemented to ensure compatibility with real-time applications. To determine the winner of the challenge, submissions were ranked based on these metrics, including runtime per frame, and placed on a leaderboard using a rank-then-mean scheme. The evaluation is described in detail in Section 2.4.

All relevant source code created in the context of the challenge organization was made available under an open source license. This included scripts for data preparation, the evaluation code, a baseline algorithm, and code for the final analysis. https://github.com/LMUK-RADONC-PHYS-RES/trackrad2025.

The top 5 participants were also required to make their methods and source code publicly available, allowing future development and research. Participants were allowed to use publicly available additional datasets and pre-trained models, as long as they were made publicly available before the start of the challenge.

2.2. Challenge timeline

The challenge was announced in December 2024 and consisted of four phases: the training and development phase, a preliminary testing phase, the final testing phase, and a post-challenge phase.

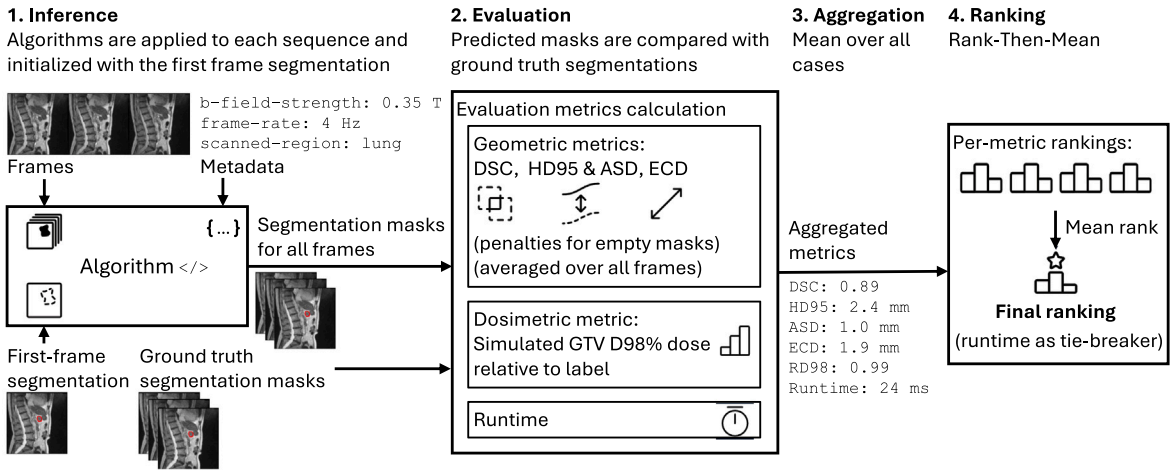

Fig. 1. Illustration of the TrackRAD2025 challenge setup. (1.) Submitted algorithms are executed using cine-MRI frames, the ground truth label segmentation mask of the first frame, and some metadata. (2.) The resulting predicted segmentation masks are evaluated in comparison to the ground truth labels. (3.) The evaluation metrics are aggregated for all frames and all cases/patients. (4.) The aggregated metrics are then used to determine a ranking, employing a Rank-Then-Mean scheme.

The training and development phase commenced with the release of the training data on March 15, 2025. The dataset included both labeled and unlabeled data (see Section 2.3), enabling participants to explore and evaluate algorithms using supervised and unsupervised training approaches. To raise awareness and encourage participation, the challenge was promoted at ESTRO 2025 on May 4, 2025 (Landry et al., 2025b).

The preliminary testing phase began on June 1, 2025. During this phase, teams could submit up to 10 proposals, which were evaluated based on 8 cases from the preliminary testing dataset. This phase aimed to familiarize participants with the challenge platform and included open logs for debugging purposes. The results were made public on a leaderboard to encourage continued development. Although the ground truth segmentation masks were concealed to prevent biased results, open logs theoretically made them vulnerable to extraction using malicious code. However, no such activity was observed.

The final testing phase started on July 17, 2025. In this phase, participants could submit algorithms for evaluation on 50 cases, the final testing dataset. To minimize overfitting to the test data, each team was only allowed two submissions, with only the better one considered for the final leaderboard. This two-submission limit allowed teams to submit different variants of their algorithms or a second version that was further improved over the course of the final testing period. To ensure the integrity of the evaluation, both the test data and ground truth labels were kept entirely hidden, as the logs were not accessible to participants. The resulting metrics and leaderboard positions were also hidden from participants until the announcement of the winners.

Both the preliminary and final testing phases ended on August 15, 2025, after which no further submissions were allowed. Participants were required to submit an *Lecture Notes in Computer Science* (LNCS) format algorithm description and questionnaire by September 1, 2025. The top 5 winners were announced on September 2, 2025, and invited to present their solutions in a challenge workshop at MICCAI 2025 on September 23, 2025.

Following the competition and challenge workshop, a post-challenge phase was opened until early 2026, allowing interested participants to submit additional algorithms for evaluation on the 58 cases from the preliminary and final testing phases.

2.3. Dataset

The TrackRAD2025 challenge dataset (Wang et al., 2025) comprised 2D+t sagittal cine-MRI sequences (time-resolved images) acquired during radiation therapy treatments on 0.35 *T* (ViewRay MRIdian) and 1.5 *T* (Elekta Unity) MRI-linacs in six international centers. In total, data from 585 individual patients was included, providing over 2,878,000 unlabeled cine-MRI frames. In addition, more than 10,000 labeled frames (including more than 8000 with multiple observer annotations) were available from 108 patients. Tracking targets (typically tumors, but also surrogates in some cases) in the thorax (179), abdomen (266), and pelvis (140) were included because they can be affected by motion and reflect the most frequently treated anatomies at MRI-linacs. Detailed information on the preparation of the dataset, the characteristics of the cases, origin institutions, annotators, and the annotation process are available in the dataset paper (Wang et al., 2025).

The complete dataset was divided into three subsets: two public datasets for supervised (50 cases) and unsupervised (477) training, and one private dataset reserved for testing within preliminary (8) and final (50) testing phases. This split balanced the desire for public training and hidden testing data. Each case consisted of imaging data and metadata (anatomical region, frequency, and MRI-linac type), and, where available, ground truth labels of the tracking target.

Parts of the testing dataset were also made publicly available alongside the training dataset after the post-challenge phase closed in early 2026. Some cases will not be made publicly available due to privacy restrictions.

The distribution of field strengths in the testing data was balanced, with 27 cases from 0.35 *T* and 31 cases from 1.5 *T* MR-Linacs. At 0.35 T, the testing data included 17 abdominal, eight thoracic, and two pelvic cases, while at 1.5 T, there were 16 abdominal, nine thoracic, and six pelvic cases. For a more detailed break-down of the challenge data-sets, please refer to the corresponding data-set paper ((Wang et al., 2025), Table 2).

Ethical approval was obtained from the internal review boards or Medical Ethics committees of the data-providing institutions. The training data was released under the CC BY-NC (Creative Commons Attribution-Non-Commercial) license and made available on Hugging Face at https://huggingface.co/datasets/LMUK-RADONC-PHYS-RES/TrackRAD2025.

2.4. Evaluation

The segmentation masks produced by the algorithms submitted to the challenge platform were evaluated using four geometry-oriented metrics and one dose-related metric specific to radiotherapy. As the sixth metric, the runtime was measured to ensure that the submitted algorithms can be used in real-time, while the patient is lying in

the machine and is being treated. The overall ranking followed a Rank-Then-Mean scheme.

Finally, participants were required to adhere to additional rules as specified at (Landry et al., 2025a) and provide additional information on their methods in a form, and submit a short LNCS-format report detailing their methods.

To guide participants in their development and produce a reference level of the metrics, a baseline algorithm implementation was provided. This baseline repeats the initial label for every consecutive frame.

2.4.1. Geometric accuracy

For each cine-MRI frame, the predictions produced were evaluated by comparing the predicted mask A and the ground truth label B using image-based metrics. For cases where labels from multiple observers were available, the STAPLE (Warfield et al., 2004) mask was used. These per-frame metrics were aggregated using the mean to obtain the per-case metrics. The following metrics were used:

1. Dice similarity coefficient. The Dice similarity coefficient (DSC) between the model prediction and the ground truth is defined as:

$$\mathrm{DSC} = \frac{2|A \cap B|}{|A| + |B|}$$

2. 95th percentile Hausdorff distance. The 95th percentile of the surface distance distribution between the model prediction and the ground truth, denoted as the 95th percentile Hausdorff distance (HD95), is defined as:

$$\mathrm{HD}_{95} = \mathrm{P}_{95}\left(\max\left\{\sup_{a\in S_A} d(a, S_B), \sup_{b\in S_B} d(b, S_A)\right\}\right)$$

where S_A and S_B denote the surface points of the predicted and ground truth segmentation masks, respectively, $d(a, S_B)$ represents the shortest Euclidean distance from point a to any point in B, and similarly for $d(b, S_B)$.

3. Mean average surface distance. The mean average surface distance (MASD) between the model prediction and the ground truth is defined as:

$$\mathrm{MASD} = \frac{1}{|S_A| + |S_B|}\left(\sum_{a\in S_A} d(a, S_B) + \sum_{b\in S_B} d(b, S_A)\right)$$

4. Euclidean center distance. The Euclidean center distance between the centroids (centers of mass) of the model prediction and the ground truth is defined as:

$$\mathrm{ECD} = \|c_A - c_B\|$$

Where c_A and c_B are the centroid coordinates of the predicted and ground truth segmentations, respectively.

In cases where an algorithm produced an empty segmentation mask on a given frame with a ground truth label available, the per-frame DSC is considered 0, and the HD95/MASD/ECD are calculated as the image size along the largest dimension in millimeters.

2.4.2. Radiotherapy metrics

For each cine-MRI sequence, the predicted segmentation masks were evaluated using a surrogate dose metric, as described in Lombardo et al. (2024b). This metric estimated the accuracy of radiotherapy dose delivery using multileaf collimator tracking based on the predicted target segmentation masks.

To compute the dose metric, the ground truth label of the first frame is used to compute an approximate radiation therapy dose distribution. First a 3 mm 2D expansion is applied to the gross tumor volume (GTV) indicated by the label to obtain a planning target volume. Subsequently, the expanded mask is smoothed using a Gaussian filter with a standard deviation of 6 mm for lung and 4 mm for all other anatomical regions. This simulates dose fall-off rates similar to those observed in clinical dose distributions for lung and higher-density tissue. For each frame, the surrogate dose distribution was then shifted by the distance between the ground truth centroid position of the tracking target and the predicted centroid position. These shifted distributions were averaged to obtain a centroid-error-shifted dose. The ratio between the GTV D98% (from the cumulative dose volume histogram) for the final shifted dose distribution and the ground truth dose distribution was calculated.

$$\mathrm{DoseMetric} = \mathrm{GTV\,D}_{98\%,\mathrm{Prediction}} / \mathrm{GTV\,D}_{98\%,\mathrm{GroundTruth}}$$

Thus, a metric of 1.0 indicates no dosimetric difference between the predicted segmentation masks and the ground truth, meaning the delivery of an equivalent D98% dose, given the model's assumptions. Lower values would indicate a reduction in the delivered dose due to tracking errors.

2.4.3. Runtime

The average runtime per frame was calculated under the assumption that faster algorithms enable real-time applications at higher frequencies and are generally less computationally expensive, making them preferable.

For this purpose, the total runtime of each submission algorithm was measured for each case. The average runtime per frame was determined using linear regression of the runtime for a given case against the number of frames present in this case, to exclude the runtime attributed to initialization procedures. The average runtime per frame was also selected as a tie-breaker if necessary.

Additionally, a maximum runtime was imposed per case. Algorithms that exceeded a runtime of 1 s per frame on the provided hardware were excluded from the challenge due to concerns regarding the real-time applicability of the algorithm.

The limit was established by benchmarking a relatively large transformer model, SAM2 (small) (Ravi et al., 2024). Without model compilation or other optimizations, this model met the real-time requirement of 8 Hz on an RTX A6000 GPU from 2020, achieving a runtime of approximately 109 ms. The runtime of this model was then measured on the Grand Challenge platform to be approximately 950 ms, which is about 9 times slower. The runtime limit was therefore set to 1 s per frame, with an additional margin for container, algorithm, and data loading.

Notably, the platform advises against using the runtime for ranking purposes, as it may improve performance through hardware or software changes that are not known to challenge organizers or participants. However, all submissions to the final testing phase occurred in a very short time span. Furthermore, the challenge organizers confirmed that no significant performance change occurred during this time period.

2.5. Supporting documents

In addition to their algorithm, participants were required to submit a form detailing their method and a report in LNCS format outlining their approach to the challenge. The form is included in the supplementary material. The LNCS format reports of the top five submissions were peer-reviewed by the organizers and made available on the leaderboard. Additionally, the top five submissions were required to be publicly available as open source.

2.6. Prize eligibility

Submissions not in accordance with the additional requirements, such as the submission of supporting documents, were excluded from price eligibility. Submissions that were not eligible for prizes were still included in the evaluation, wherever possible. Submissions from organizers or related parties (determined by recent common publications) were also excluded from prize eligibility, as defined in the challenge design document (Landry et al., 2025a). Cash prizes were awarded to the top five teams (2500 USD from sponsor Elekta (Stockholm, Sweden), 1st place 1000 USD, 2nd place 600 USD, 3rd place 400 USD, 4th place 300 USD, 5th place 200 USD).

2.7. Ranking

The metrics calculated for each submission were made public on a leaderboard. Submissions were sorted by and ranked in each of the six metrics. In a Rank-Then-Mean scheme, the mean rank per submission (averaging over the six metrics) was computed to produce a final mean rank score, which was used to determine the final position in the leaderboard. Submissions that did not improve above the baseline algorithm in any of the four geometry metrics or the dose metrics were excluded from the ranking. The per-frame runtime was considered as a tie-breaker for determining the top five submissions. However, this scenario did not materialize. For the preliminary testing phase, participants were allowed up to ten submissions, which were immediately visible on the public leaderboard. For the final testing phase, participants were allowed up to two submissions, and only the highest-scoring submission was visible in the leaderboard.

2.8. Qualitative analysis

The information from the submission descriptions and forms was used to qualitatively evaluate the different approaches participants used.

2.9. Identification of failure modes

Challenging cases were identified by comparing the mean DSC for each patient case with the corresponding mean DSC over all cases from the same anatomical region and submission. Cases were considered challenging when all of the top 5 submissions exhibited a decrease in their DSC of at least 0.05 relative to their mean DSC score per anatomical region (ΔDSC). These cases were then further analyzed by visual inspection to identify the failure modes of the submissions.

2.10. Inter-observer variability

To establish an overall success criterion for model performance, the degree of agreement between human annotators was quantified. The inter-observer variability was computed two-fold: (1) pairwise between all observers, and (2) against a consensus reference obtained using the STAPLE algorithm (Warfield et al., 2004). This analysis was performed on the subset (N = 20) of the test dataset annotated independently by five observers each. The resulting inter-observer variability provides an empirical upper bound for model performance, indicating the inherent uncertainty in the manual annotations.

2.11. Quantitative analysis

For further analysis beyond the challenge ranking metrics, failure rates were calculated, defined as the percentage of frames with ECD greater than 3 mm. Additionally, the per-frame J& F metrics (Perazzi et al., 2016) were computed. The J-metric is also known as intersection over union or Jaccard value, and the F-metrics is also known as F2-score, a metric combining recall and precision. The correlation coefficient (Spearman, 1904) of these metrics was computed. A minimal subset of relevant metrics was identified.

2.12. Group-based analysis

An impact analysis was conducted to evaluate how the anatomical region (thorax, abdomen, or pelvis) and magnetic field strength (0.35 T or 1.5 T) affect the performance metrics of the top five submissions For each submission, the statistical significance of the differences in the DSC between the two field strengths was assessed using a Mann–Whitney U test (Mann and Whitney, 1947). Additionally, discrepancies among the three anatomical regions were examined using a Kruskal–Wallis test (Kruskal and Wallis, 1952). A significance level of $\alpha = 0.05$ was applied in both tests. Where necessary, a post hoc Dunn test (Dunn, 1964) was performed to identify discrepancies among the three anatomical regions groups in pairwise comparisons.

2.13. Ranking stability

The stability of the ranking was evaluated, following (Wiesenfarth et al., 2021; Huijben et al., 2024). First, the variability of the rankings with respect to the cases included in the test set was assessed. For this purpose, bootstrapping was used to generate 1000 test sets, each consisting of 50 randomly selected cases from the test dataset, with cases potentially being selected more than once. The variability between the original ranking and the rankings from the 1000 bootstrapping test sets was assessed using Kendall's Tau analysis (Kendall, 1938). Additionally, the stability of the rankings was assessed in relation to variations in the evaluation metrics. For this purpose, rankings were recomputed using different sets of metrics: the official set of metrics with each metric removed in turn (ablation analysis), a minimal subset of metrics identified based on the correlation analysis, the full extended set with the failure rate and J&F metrics, or with the runtime metric weighted double or triple, and the official set of metrics without applying penalty terms for empty predictions (by repeating the last-non-empty prediction). This analysis enabled an assessment of whether the specific choice of evaluation scheme had a meaningful impact on the resulting rankings, i.e., on the outcome of the challenge.

2.14. Dataset variability

The SAM-2 foundation model (SAM2-large variant) was used to assess the consistency of the annotations within each scan quantitatively and to identify outlier frames. For each sequence from the final testing set, SAM-2 was prompted with every frame and its corresponding ground truth label to segment the remaining frames. The DSC metric was computed between the predicted segmentation masks and the reference labels, and the results were assembled into a pairwise consistency matrix. Frames with consistently low DSCs across comparisons below a threshold of Q1 - 1.5 · IQR were classified as outliers, indicating potential labeling inconsistencies or atypical image appearances. These detected outliers were excluded to generate a cleaned test set. The metrics for submissions were then recalculated with and without the outlier frames to evaluate the robustness of the performance rankings. This comparison enabled an assessment of whether the detected outliers affected the relative performance of the models.

3. Participation

The TrackRAD2025 challenge received notable participation and interest from the target demographic. By the end of the final testing phase, 148 participants from a total of 28 countries registered for the challenge on the challenge platform.

In total, 166 algorithms were submitted to the preliminary test phase, resulting in 100 successful submissions. 29 algorithms were submitted to the final testing phase, resulting in 24 successful submissions. These belonged to 15 unique teams. Ten teams completed the additional prize eligibility requirements and submitted both the required form and LNCS report. Two teams submitted the form, but no reports were provided. The submissions from one team were excluded from the leaderboard because they did not improve on the baseline algorithm in any of the four geometric metrics or the dose metric. These submissions reproduced the baseline. Table 1 shows an overview of the algorithms submitted.

4. Submissions

The authors of the top five submissions were invited to present their methods in the following section in more detail. In addition, the authors of some technically interesting submissions were also invited. The LNCS reports for all of these submissions are hosted on the challenge platform along the leaderboard and contain references on how to make use of the respective methods. They can also be found in the supplementary material.

Table 1
Ranking, Team or submission name, minimal submission descriptions of the successful participants. Abbreviations used in the description: finetuned (FT) and with pseudolabels (wPL). Italized descriptions are based on submission forms only, not on complete LNCS reports.

Rank	Team name	Description
1	Track'n'Treat	MedSAM2 FT wPL
2	Hkini	SAM2 FT wPL
3	CoTrackRAD	Contour point tracking
4	CIT	SAM2 FT w adapted loss
5	Reg'n'Track	SAM2 FT + ensemble
6	Breizhtrack	SAM2 FT
7	Deeptrack	C-VMorph; registration
8	Youbetcha	*SAM2 FT + nnUNet*
9	Biomed	TransMorph + Attention
10	hnourzad	*-/no report*
11	Peter Treloar	Cross correlation techniques
12	ShangxuanLi	*-/no report*
13	jintao	*SAM2 FT wPL*
14	lWM	*-/no report*

4.1. Top 1 - track'n'treat - label-efficient semi-supervised cine-MRI tumor tracking using MedSAM2

The winning submission developed two MedSAM2-based strategies (Ma et al., 2025). The first approach (Method A) finetuned MedSAM2 on 40 labeled cases from the training set, using the initial frame ground truth label to derive a tight bounding box prompt to guide segmentation across subsequent frames. To improve generalization without added inference cost, the authors applied a lightweight "model soup": a rank-weighted parameter average of the best validation checkpoint and its two nearest neighbors, yielding a single robust network.

The second approach (Method B) introduced a semi-supervised ensemble. Two MedSAM2 models were trained on different labeled splits and augmented with curated pseudo-labels generated by scanner-specific nnU-Net models (Isensee et al., 2021). Reliability of pseudo-labels was enforced via adaptive frame quality filtering, uncertainty-based acceptance, and intra-sequence consistency checks, followed by visual quality control. At inference, the two MedSAM2 models are combined by probability-level averaging to produce the final segmentation mask.

Both methods were scanner-agnostic, operating across 0.35 T and 1.5 T data. As the final leaderboard reports only each team's best submission, the listed result corresponds to Method A. In a comprehensive ranking that includes all submissions (see Table S1 in the supplementary material), Method B would place second.

4.2. Top 2 - Hkini - a three-stage semi-supervised finetuning method for SAM2 in real-time MRI tumor segmentation

The submission adapted SAM2 for tumor tracking through a progressive three-stage semi-supervised finetuning pipeline. In stage one, several SAM2 variants were evaluated on a held-out validation set, and the lightweight sam2_hiera_tiny model was selected for its balance between accuracy and real-time efficiency. In stage two, full-parameter supervised finetuning was performed on 38 sequences from the labeled training dataset, using first-frame prompts (mask, bounding box, and center point) to specialize SAM2 for cine-MRI tumor segmentation. This stage yielded notable improvements in segmentation accuracy and spatial consistency. Stage three introduced semi-supervised learning by generating pseudo-labels for 100 unlabeled sequences with the stage two model. These pseudo-labels, obtained with bounding box prompts, were combined with the labeled data to form an expanded training set of 138 sequences. The model was then finetuned again, with adjusted loss weights and learning rates to reduce pseudo-label noise and improve robustness across centers, anatomical sites, and MRI field strengths.

4.3. Top 3 - CoTrackRAD - tumor tracking as point tracking: CoTracker3 in the TrackRAD2025 challenge

The submission reframed the task of tumor tracking as a boundary point tracking task, leveraging the pre-trained CoTracker3 model initially developed for natural video analysis (Karaev et al., 2023).

Instead of propagating full segmentation masks, the tumor contour provided in the initialization frame was converted into a discrete set of boundary points. These points were then tracked across time using CoTracker3, which produced temporally coherent trajectories. The tracked boundary points were then converted back into segmentation masks at each time step to produce full-frame tumor delineation. The CoTracker3 model checkpoint used was trained exclusively on large-scale natural and synthetic video data and transfers to cine-MRI without any additional training. This approach therefore did not require any medical data nor annotations.

4.4. Top 4 - CIT - An adapted SAM2.1 framework for real-time tumor segmentation in MRI-guided radiotherapy

This method adapted the Segment Anything Model 2.1 (SAM2.1) base-plus variant. Since SAM2.1 was initially designed for RGB images, the vision input layer was modified for single-channel grayscale cine-MRI by averaging its weights across channels. Training primarily followed SAM2.1's default configuration, utilizing the AdamW optimizer, a base learning rate of 5e-6, a learning rate of 3e-6 for the vision encoder, and a cosine decay scheduler. To improve tumor-specific tracking performance, the standard SAM2.1 loss functions (focal, Dice, mean absolute error, and cross-entropy) were extended with five additional domain-tailored losses: boundary, center distance, surface, temporal, and motion. Preprocessing employed bounding-box cropping with min–max and Z-score normalization, while data augmentation included random flips, rotations, zooms, and crops. The model was finetuned for 250 epochs, and the best checkpoint was selected based on the highest average validation score. During inference, frame intensities were clipped relative to the initial frame's distribution, and a Gaussian smoothing filter was applied to mask boundaries for stability.

4.5. Top 5 - Reg'n'track - No chains, just gains: Hierarchical SAM2 motion-tracking

This method used Segment Anything Model 2.1 (SAM2.1) with a hierarchical motion-tracking framework (HIM2SAM) (Chen et al., 2025). The approach leveraged the initial tumor mask provided in the first frame to guide the propagation of segmentation across time-resolved cine-MRI sequences. HIM2SAM exploited both temporal consistency and spatial embeddings from SAM2.1 to maintain accurate and stable tracking. Three SAM2.1 variants (tiny, small, and base+) were finetuned using the TrackRAD2025 training data. To improve robustness, their frame-wise logits were selected and assembled using a motion estimation system to produce temporally coherent segmentation masks. Notably, the sequential inference of multiple variants for each timestep resulted in a longer runtime compared to the other SAM2-based methods.

The pipeline was designed to be anatomy-agnostic and applied uniformly across all anatomical regions and MRI field strengths. Preprocessing involved Z-score intensity normalization and region-of-interest extraction to focus learning on relevant structures. Postprocessing steps included soft-label averaging of predictions and morphological refinements to smooth and stabilize boundaries.

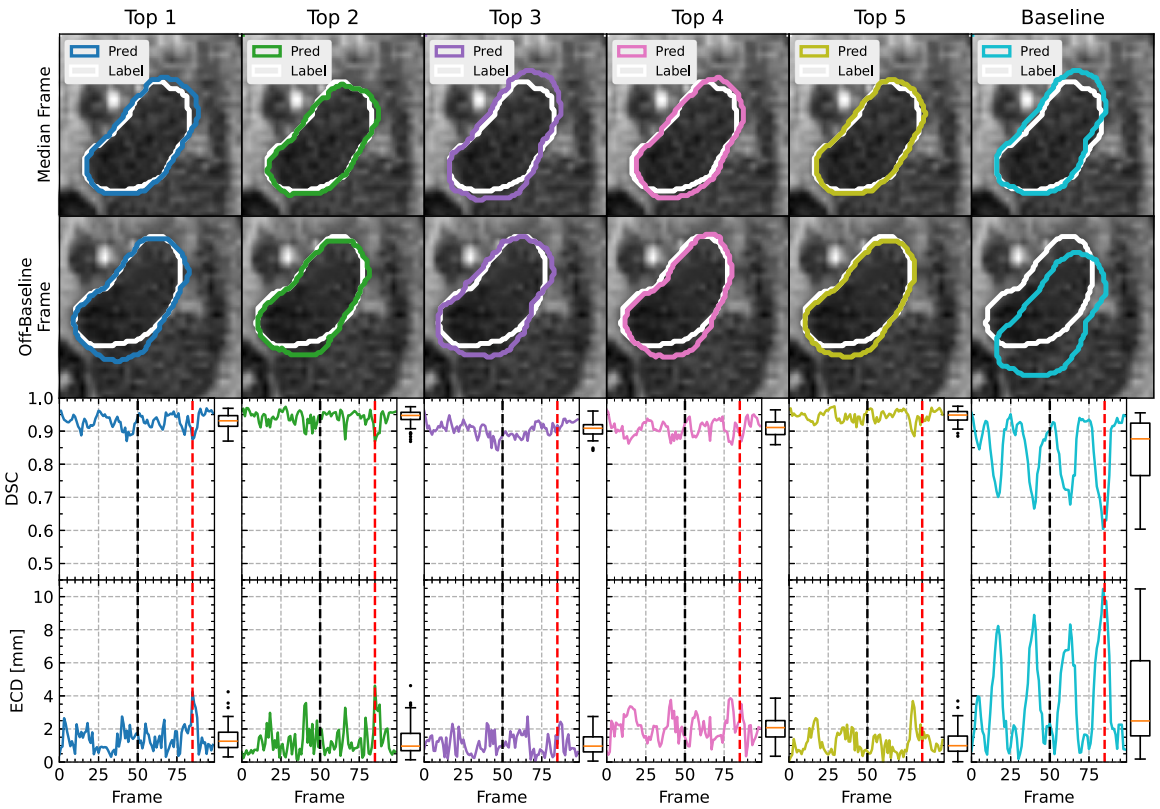

Fig. 2. Example target tracking results. Each column corresponds to a top 5 model or the static registration baseline. The top two rows show the segmentation mask for two frames, where white contours denote labels and colored contours indicate model outputs. The bottom two rows present the DSC and ECD metrics over the frame index. Vertical dashed lines mark the median (black) and off-baseline (red) frames shown above. Adjacent boxplots summarize the distributions of metrics across all frames.

4.6. Honorable mention - Breizhtrack - Finetuning segment anything for accurate tumor tracking in MRI sequences

This submission evaluated two complementary strategies: an unsupervised registration approach based on IMPACT (Boussot et al., 2025) and several foundation-model-based segmentation methods, including SAM2.1 and its variants. After comparing these models on the annotated dataset, the SAM2-based approach was selected to satisfy the strict one-second runtime constraint and was subsequently finetuned to further improve performance. The SAM2.1b+ checkpoint was finetuned on the labeled portion of the training data, without a separate validation split. A unified inference strategy was applied across all MRI types, and test-time augmentation was tested but found unnecessary.

4.7. Honorable mention - Deeptrack - cascaded U-net for real-time deformable registration in MRI-guided tumor tracking

This submission proposed a registration-based approach for real-time tumor tracking in cine-MRI, introducing a VoxelMorph network named C-VMorph. The model was designed to progressively refine displacement vector fields (DVFs) by applying the same module for four iterations, aligning moving cine-MRI frames with a fixed target to capture tumor motion. Its cascaded U-Net architecture incorporated ROI-guided supervision, ensuring that training emphasized relevant regions. The model was trained in two phases. First, using only the unlabeled dataset, the model was trained to minimize a mean squared error loss for image alignment and a DVF regularization term based on bending energy, to encourage smooth motion fields. In a second supervised phase, the labeled data was used to minimize a DSC-based loss as well. A small part of the labeled dataset was used for validation.

4.8. Honorable mention - Biomed - attention-guided TransMorph for real-time tumor tracking in cine-MRI

The proposed method was based on deep learning-based deformable image registration, improving the TransMorph architecture (Chen et al., 2022). It followed a two-step training paradigm. First, unsupervised pretraining was performed on unlabeled image pairs to learn general alignment patterns without requiring manual labels. Second, supervised finetuning used segmentation labels to refine the model for accurate tumor localization. To enhance focus on relevant regions, attention gates were integrated into skip connections, enabling the model to prioritize important anatomical features during registration. The optimization employed a compound loss function that combines boundary-weighted Dice loss for segmentation accuracy, adaptive mean squared error (MSE) for alignment, L1 loss for stability, smooth diffusion for deformation regularization, and edge-based regularization to preserve structural details. This yielded significant performance improvements over the baseline TransMorph model, demonstrating enhanced segmentation accuracy, deformation quality, and robustness for real-time tumor tracking.

5. Results

All successful submissions tracked the defined targets across cine-MRI frames. As depicted in Fig. 2 for one case from the final testing set, the top 5 models consistently produce good segmentation masks, compared to the no-registration baseline method, which results in degraded metrics due to motion.

Table 2
Evaluation metrics and leaderboard of the ranked submissions to the final testing phase. Submissions are ordered by mean rank across all evaluation metrics (Rank-Then-Mean): Dice similarity coefficient (DSC, higher is better), mean average surface distance (MASD, mm, lower is better), 95th percentile Hausdorff distance (HD95, mm, lower is better), Euclidean center distance (ECD, mm, lower is better), relative D98 dose coverage (higher is better), and runtime per frame (TPF, ms, lower is better). Additionally, the metrics from the static registration baseline are presented, along with the inter-observer metrics (IO), both pairwise and against a STAPLE mask.

Rank	Team	DSC ↑	MASD (mm) ↓	HD95 (mm) ↓	ECD (mm) ↓	Relative D98 ↑	TPF (ms)↓	Mean Rank
1	Track'n'Treat	0.891 (1)	1.5 (1)	4.2 (1)	1.7 (1)	0.936 (4)	43 (7)	2.5
2	Hkini	0.886 (2)	1.5 (2)	4.5 (2)	1.9 (2)	0.963 (1)	61 (11)	3.3
3	CoTrackRAD 3	0.876 (5)	1.7 (3)	4.5 (3)	1.9 (3)	0.925 (7)	18 (2)	3.8
4	CIT	0.871 (6)	2.0 (4)	5.3 (4)	2.2 (4)	0.939 (2)	24 (4)	4.0
5	Reg'n'Track	0.880 (3)	2.1 (5)	5.6 (7)	2.2 (5)	0.932 (5)	515 (14)	6.5
6	Breizhtrack	0.879 (4)	2.3 (7)	5.6 (6)	2.6 (8)	0.936 (3)	75 (12)	6.7
7	Deeptrack	0.851 (9)	2.3 (6)	5.3 (5)	2.3 (6)	0.908 (9)	55 (9)	7.3
8	Youbetcha	0.863 (8)	2.5 (8)	6.3 (10)	2.8 (9)	0.930 (6)	45 (8)	8.2
9	Biomed	0.836 (13)	2.6 (9)	6.2 (9)	2.5 (7)	0.875 (13)	22 (3)	9.0
10	hnourzad	0.864 (7)	11.8 (13)	14.6 (13)	12.0 (13)	0.912 (8)	18 (1)	9.2
11	Peter Treloar	0.842 (12)	3.1 (10)	5.9 (8)	2.8 (10)	0.899 (11)	225 (13)	10.7
12	ShangxuanLi	0.846 (11)	4.3 (11)	7.8 (11)	4.1 (11)	0.896 (12)	57 (10)	11.0
13	jintao	0.849 (10)	15.8 (14)	18.6 (14)	16.2 (14)	0.905 (10)	41 (6)	11.3
14	lWM	0.809 (14)	5.3 (12)	10.5 (12)	4.9 (12)	0.701 (14)	27 (5)	11.5
	Baseline	0.773	5.7	8.1	3.6	0.736		
	IO (pairwise)	0.853	0.65	6.50	4.0			
	IO (staple)	0.891	0.13	4.2	2.7	0.900		

5.1. Metrics

As shown in Table 2, the evaluation on the challenge platform demonstrated that the submissions achieved high-quality target tracking on the final testing dataset. The winning approach, Treat'n'Track, achieved mean and standard deviation evaluation metrics of DSC 0.89 ± 0.05, HD95 4.2 ± 1.8 mm, and ECD 1.5 ± 0.99 mm at a runtime of 42 ms per frame. The top five methods showed mean DSC >0.87 and ECD <2.1 mm. The mean pairwise inter-observer variability was DSC 0.853, and variability against the STAPLE mask was DSC 0.891, also included in Table 2. The segmentation accuracy of the top 5 submissions was thus comparable to inter-observer variability. All top 5 models, except one, were able to achieve a fast runtime on the provided hardware, with a computing time per frame below 125 ms, and would thus have been fast enough for real-time target tracking at 8 Hz. The exception to this was the top 5 submission Reg'n'Track, which employed an ensemble approach, combining the outputs and runtimes of multiple SAM2 variants.

The penalty for empty frames was applied sparingly. It primarily affected two submissions: the Top 10 (Hourzad) submission, which had a total of 104 empty predictions distributed across four affected cases, and the Top 13 (Jintao) submission, which had 112 empty predictions distributed across four affected cases. In addition, four other submissions produced between 1 and 4 empty predictions: Top 5 and Top 6 (2 each), Top 12 (4), Top 7 (1).

5.2. Group-based analysis

The impact of magnetic field strength and anatomical region on model performance was assessed. Compared by magnetic field strength (Fig. 3 a), no significant differences in performance were observed between the 0.35 T and 1.5 T cases for most models. Only for the Top 3 model there was a statistically significant but small difference, with 0.35 T cases producing higher DSC scores compared to 1.5 T. In contrast, the anatomical region was found to exert a more substantial influence on the segmentation performance (Fig. 3 b). Specifically, pelvic cases consistently produced significantly higher DSC values in all top submissions and the baseline algorithm, compared to thorax and abdomen cases. Between these, there were no statistically significant differences. These results matched expectations, as both magnetic field strengths are generally suitable for MRIgRT and exhibit sufficient contrast for accurate target localization. Additionally, the anatomical region of the pelvis exhibits less motion, and therefore, target tracking solutions should generally yield better results. For submissions using SAM2-based methods and those using other methods (Fig. 3 c), SAM2-based submissions consistently achieved a higher mean DSC score.

5.3. Qualitative analysis

A qualitative analysis of the submissions revealed that participants extensively utilized foundation models. Of the eleven teams that submitted forms and/or reports, six teams used the SAM2 foundation model for video and image segmentation. The winning submission also utilized this model, specifically the MedSAM2 variant, which is a variant of SAM2 finetuned on a large dataset of medical imaging data. Most teams conducted additional training to optimize the SAM2 model further.

All teams that implemented training utilized the 50 labeled cases provided in the TrackRAD2025 dataset for training purposes. The two top submissions both performed SAM2-based pseudo-labeling of the unlabeled dataset to obtain additional training cases. The training itself generally followed the training setup of SAM2 itself. One team (top 4) adapted the loss function to reflect the evaluation metrics of the TrackRAD2025 challenge. Two teams further modified the training procedure to combine multiple checkpoints into a model soup (top 1) or multiple models into an ensemble (Top 5 and Top 1 - unranked method B). One submission combined the finetuned SAM2 with a specifically trained nnUNet to further improve the SAM2 outputs. The 3rd place submission did not use SAM2 but CoTracker3, a foundation model for point tracking, reformulating the task of video object segmentation as tracking the points on the boundary of an object mask. For this purpose, the CoTracker3 model was used without modifications or further training. Two other teams (Top 7 & 9) used models performing image registration. Both models could be trained without using labels and utilized both unlabeled and labeled training data, with the labeled data partially used for validation or additional supervised training. The Top 7 submission trained CNN-based models for image registration, with individual models for each B-field strength and anatomical region. The Top 9 submission added an attention mechanism to the TransMorph architecture. Finally, one team (Top 11) submitted an algorithm based on classical/non-machine learning, using cross-correlation techniques. Only one team utilized the provided metadata, with the top 7 submissions training dedicated models for each MRI-linac type and anatomical region. No team utilized the provided frame rate.

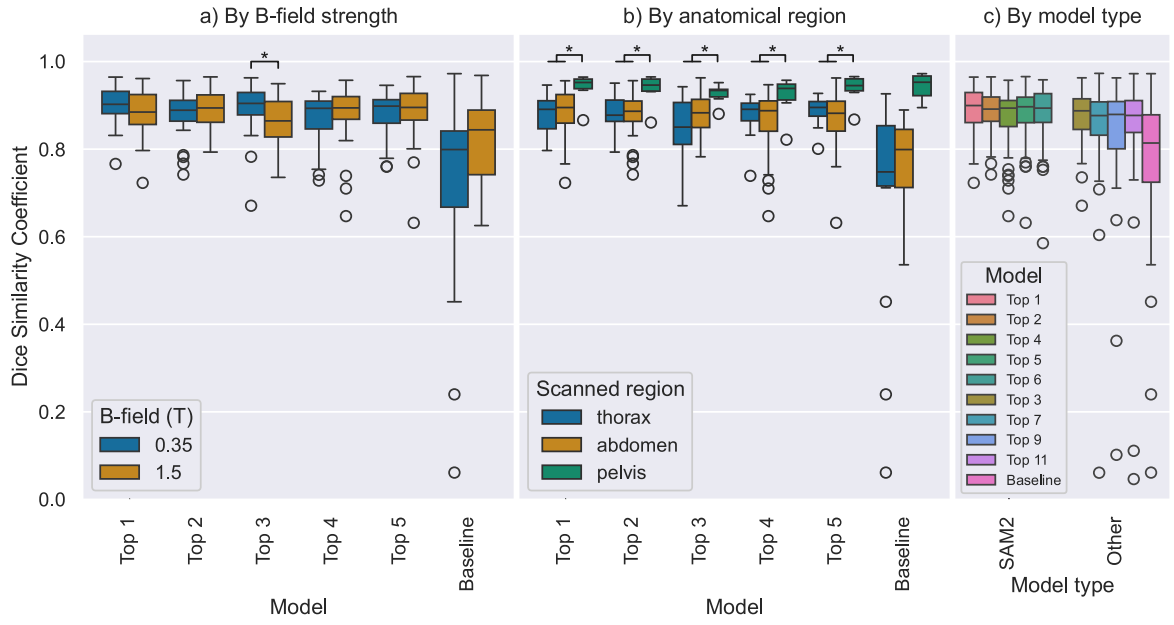

Fig. 3. Visualization of the performance signaled by DSC of the top 5 submissions for cases grouped by (a) different B-field strengths and (b) different anatomical regions. The baseline algorithm is included to indicate the general presence of movement. Panel (c) shows the performance between model types for SAM2-based and for other submissions. 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. Brackets marked with a star indicate statistically significant differences between the different case groups.

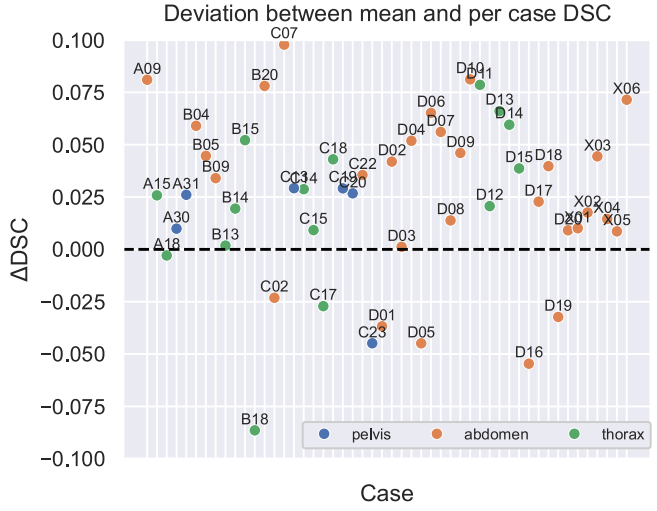

Fig. 4. Deviation of a case's mean DSC from the mean DSC over all cases. The mean DSC over all cases was calculated per submission and per anatomical region. The submission with the best performance per case is plotted for better visibility, indicating that the remaining 4 submissions out of the top 5 had lower ΔDSC. Case IDs are printed next to the data points. Only the cases B18 and D16 exhibit a decrease of at least 0.05 in the DSC metric for all top 5 submissions.

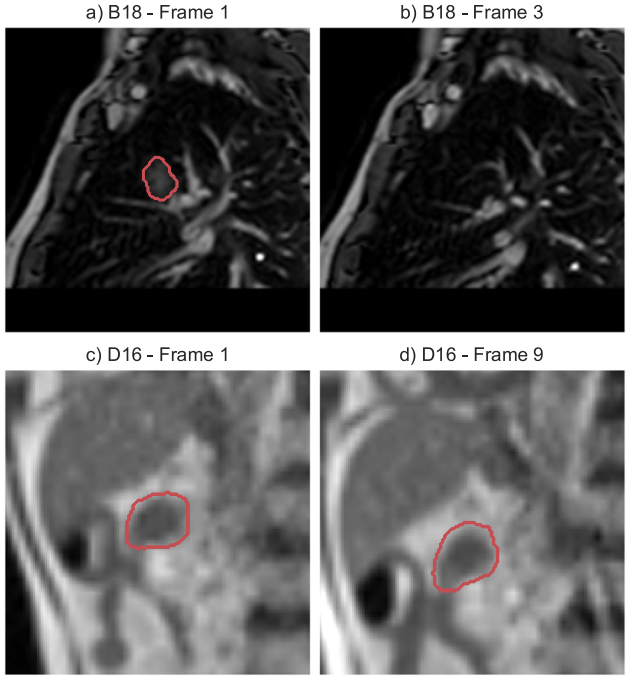

Fig. 5. Examples of the two challenging cases identified as per the ΔDSC definition from Section 2.9.

5.4. Identification of challenging cases and failure modes

Only two cases, B18 and D16, were identified as particularly challenging according to the ΔDSC definition of Section 2.9, with all top 5 submissions having at least 0.05 worse DSC than the mean DSC over cases from the same anatomical region (Fig. 4).

Detailed analysis of these cases revealed two distinct failure modes. In case B18 (Fig. 5a–b), motion perpendicular to the imaging plane resulted in the target being partially or completely absent in several frames. In case D16 (Fig. 5c–d), the target-tissue interface was particularly blurry. Detailed analysis of the other cases with a DSC worse than the mean DSC over cases from the same anatomical region showed similar features, although less pronounced.

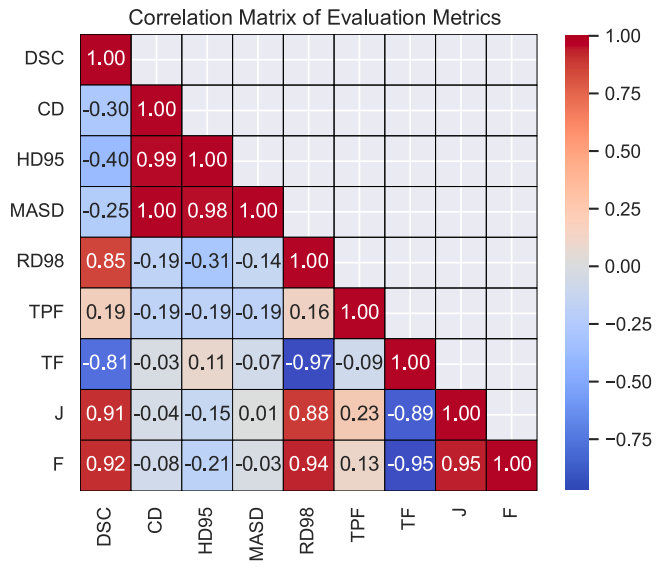

Fig. 6. Correlation of the different evaluation metrics used (Dice similarity coefficient (DSC), Center Distance (ECD), 95th percentile Hausdorff distance (HD95), Mean average surface distance (MASD), Relative D98 dose metric (RD98), and Runtime per Frame (TPF)) and additionally considered evaluation metrics (Tracking failure rate (TPF) and J&F metrics).

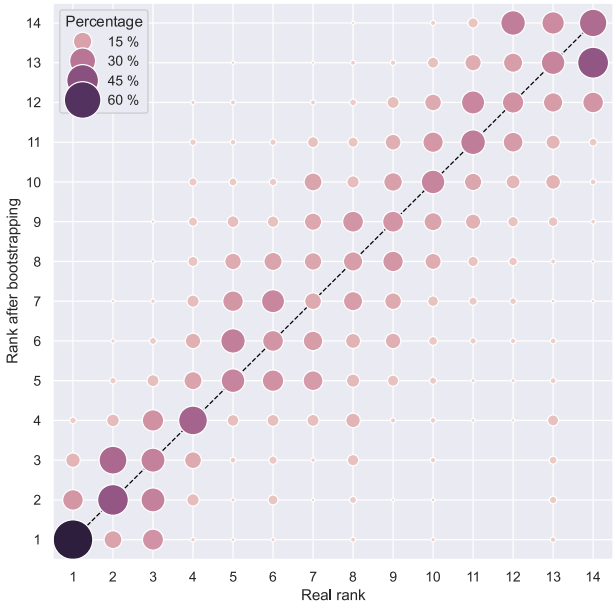

Fig. 7. Visualization of ranking stability with respect to case selection. Circle size and color are proportional to the frequency of the rank achieved based on the respective submission on N = 1000 bootstrapped sets of 50 cases (with repetitions). A dashed line indicates the identity, i.e., the location from the ranking on the full dataset.

5.5. Metrics correlation

The correlation between the metrics was calculated (Fig. 6) for the aggregated metrics of all submissions on the leaderboard. The distance metrics, ECD, HD95, and MASD, showed a very high correlation coefficient above 0.98 with each other. The DSC showed a weak negative correlation with these metrics, with correlation coefficients of −0.3, −0.4, and −0.25, respectively. The radiotherapy-specific dose metric showed a strong correlation with the DSC metric (r = 0.85) and a weak correlation with distance metrics of −0.19, −0.31, and −0.14 for ECD, HD95, and MASD, respectively. A longer runtime was weakly correlated with better evaluation metrics, i.e., higher DSC and RD98 and lower ECD, HD95, and MASD. The additionally considered tracking failure rate and the J and F metrics were highly correlated with the DSC and RD98 metrics. In general, these correlations indicate that the selected metrics were sufficient to evaluate various aspects of submission performance. DSC, ECD, and runtime per frame form a minimal subset of metrics that reduce correlations.

5.6. Ranking stability

A variability analysis of the results was conducted to determine if the rankings of the challenge remain consistent under modifications to the challenge case selection, the chosen metrics, the penalty for empty predictions, and the consistency of the ground truth labels.

5.6.1. Case variability

Analysis of the variability of the rankings with respect to the cases included in the test set (Fig. 7) showed that the ranking was robust with respect to the case selection. The official ranking overall matched the hypothetical rankings based on bootstrapped test sets (N = 1000), with Kendall's Tau correlation coefficients for the real and the bootstrapped rankings showing a high average correlation coefficient of $\tau = 0.749$ with $p = 0.003$.

5.6.2. Metrics variability

Hypothetical rankings were generated using different sets of metrics, including the default set with single metrics removed, a minimal set of metrics (DSC, ECD, and TPF), and the default set extended with the tracking failure rate or the J and F metrics, or with the runtime metric weighted double and triple. As depicted in Fig. 8, modulating which metrics are included has only a slight effect on the final ranking. Specifically, regarding the top 5 submissions, the top 4 are consistent between all considered sets of metrics. Furthermore, only when the runtime metric was weighted triple did the highest ranking submission change. Overall, these results show that the selected metrics provide a robust ranking.

5.6.3. Empty predictions penalty

Another possible alteration to the metrics that was discussed during challenge development was the implementation/handling of empty predictions for non-empty ground truth segmentation masks. For the final challenge, the organizers decided that empty frames should be counted as frames with a zero DSC value, with no dose contribution for the RD98 metric, and metric values for HD95, ECD, and MASD equal to the longest side of the frame. This results in even a single empty prediction severely impairing the overall metrics of a model. While participants were informed about this penalty and instructed to prevent empty predictions or to reuse the last non-empty frame if nothing else, the implementation of this penalty could still have impacted the overall leaderboard. To confirm the robustness of the challenge results with respect to this design decision, hypothetical rankings were calculated by auto-filling empty predictions with the previous non-empty prediction. As depicted in Fig. 8, applying this modification to the evaluation did not affect the ranking, except for the two submissions with the most empty predictions. These moved from tenth to third and from the 13th to eighth, respectively. The ranking of the other submissions, which had only a small number (≤ 4) of empty predictions, remained unchanged.

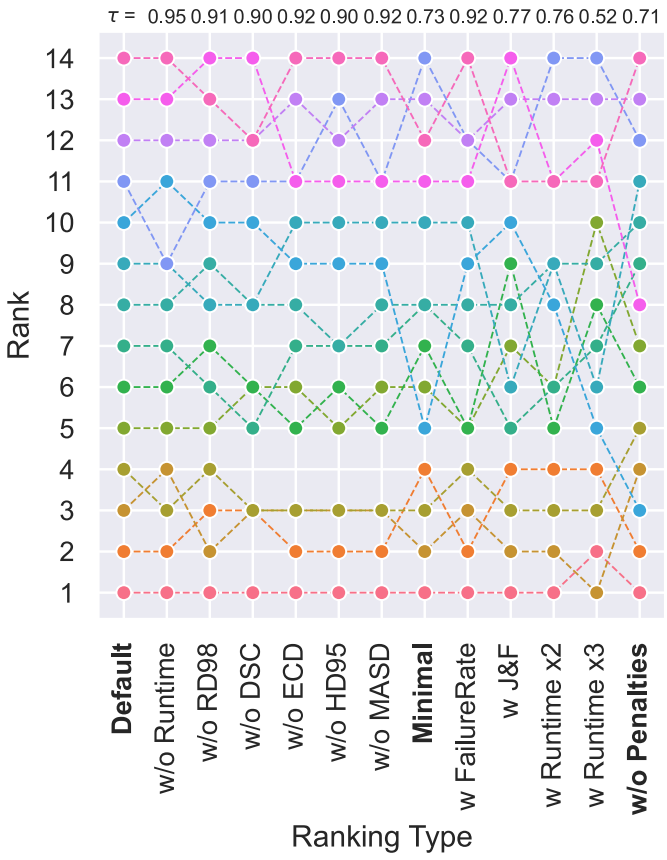

Fig. 8. Hypothetical rankings based on other sets of metrics included in the Rank-Then-Mean scheme. The actual default metrics used are displayed on the left, with each submission positioned according to its corresponding rank. Lines connect the submissions to their positions on hypothetical other sets of metrics. Kendall's Tau correlation coefficients for the real and hypothetical rankings are shown above.

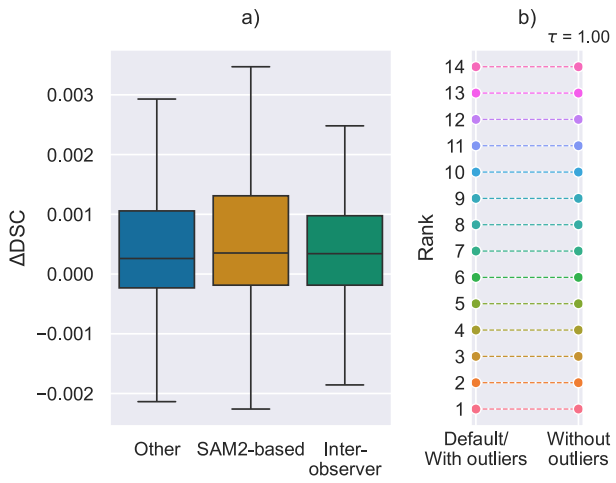

Fig. 9. Impact of detected outliers on overall per-case metrics (a) and rankings (b). In (a), the impact of the outlier frames on per-case metrics is shown aggregated for SAM2-based and other submissions, and inter-observer variability. In (b), a hypothetical ranking without the outlier frames and the actual ranking are shown, with lines connecting the position of each submission between both. Kendall's Tau correlation coefficient for the real and the hypothetical rankings is shown above.

5.7. Dataset variability

A SAM2-based outlier detection was used to identify frames for which the ground truth labels aligned less with the other labels of the same sequence. The number of such frames varied per case, ranging from 0 to 17.0% of the frames. The number of outlier frames per case was not normally distributed (Shapiro–Wilk, $p = 0.003$) with a median of 4.6% and an inter-quartile range of 2.1%–7.0%. No statistically significant differences were found in the percentage of outlier frames between cases from different institutions (ANOVA, $p = 0.819$). In total, 5.2% of the frames were identified as outliers. Excluding these outlier frames from the metric computation resulted in improved mean metrics for the submissions. On average, the DSC metric increased by 0.0012 across all cases and submissions. This indicated that the detection method indeed primarily identified frames for which the submissions produced below-average results. Comparing the impact of outlier-frame exclusion on metrics between models based on SAM2-derived architectures or others (Fig. 9 a) did not show statistically significant differences (Mann–Whitney U, $p = 0.121$). This indicated that the SAM2-based outlier detection not only excludes frames that are difficult for SAM2-based models but also those for which the predictions from all models deviate from the ground truth.

Comparing the inter-observer variability with and without the outlier frames (Fig. 9 a) showed similar results. This indicates that the detected outlier frames are indeed those in which the delineation deviates from the others in the same sequence by the same annotator. Comparing the default leaderboard ranking with all frames with a hypothetical leaderboard excluding the identified outlier frames (Fig. 9 b) showed no impact on the leaderboard. This suggests that the potential inconsistencies in the delineation detected by the SAM2-outlier detection did not significantly impact the overall ranking.

Although not affecting the ranking robustness in this evaluation, it should be noted that the use of SAM2 to detect outlier frames could still bias and artificially inflate the performance of SAM2-based and other methods by removing frames that might represent genuine anatomical challenges where the model simply fails, rather than actual labeling inconsistencies.

6. Discussion

The TrackRAD2025 challenge was designed to determine the leading methods for target tracking in 2D+t cine-MRI for MRIgRT.

6.1. Impact

The TrackRAD2025 challenge allowed for a detailed comparison of algorithms and machine learning techniques for target tracking in MRIgRT. It was the first large-scale, multi-center challenge for this task, garnering significant participation with a total of 166 participants and 124 submissions. All valid submissions to the final testing phase were generally able to produce target segmentation masks that outperformed the baseline algorithm of static registration. The top five approaches achieved mean DSC values above 0.87, HD95 below 5.6 mm, and ECD below 2.1 mm, comparable to the inter-observer variability against STAPLE (DSC: 0.89, HD95: 4.2 mm, ECD: 2.7 mm). These results indicate a high level of tracking accuracy. The results are also comparable to recent literature (Peng et al., 2024; Lombardo et al., 2024b; Blöcker et al., 2024). Analysis of the methods used by participants revealed that finetuning foundation models yielded the best results. The comparison between the submissions based on the SAM2 foundation model and those based on other methods highlighted that fine-tuned applications of SAM2 (Top 1, 2, 4, 5, and 6) consistently achieved high performance. Without finetuning, the application of CoTracker3 (Top 3) also yielded high performance, with a lower runtime per frame.

6.2. Limitations

Although the TrackRAD2025 challenge successfully achieved its goal of determining the leading MRIgRT target tracking methods, it was limited by several issues. Most importantly, the dataset used in the final testing phase to rank the submissions could have been larger and more diverse. Although the public training dataset features a reasonable number of unlabeled frames for training, it only features a significantly smaller number of labeled cases. This was due to the significant effort required by the fully manual labeling process. Although this ensured a high quality of the segmentation masks, it limited the number of cases to be included in both the public labeled training dataset and the private testing dataset. Another limitation of the challenge was the set of metrics selected for evaluation. Only a dose surrogate has been adopted to evaluate submissions, which may not accurately reflect real dose distributions. In particular, this metric also does not account for tissue deformation or proximity to organs at risk (OARs). Consequently, a high score on this metric implies target coverage, but offers no guaranty on OAR avoidance or toxicity, which is one of the main constraints in real-world clinical gating. Further, surface distance metrics were overrepresented compared to other metrics (two out of six metrics). A more diverse set of metrics could have been helpful, although the variability analysis demonstrated the overall robustness of the resulting ranking against differing metric selections. However, weighing the runtime with a factor of three would have resulted in a different highest ranked model, not based on SAM2 as foundation model, but on CoTracker3. Therefore, future iterations could place a higher weight on computational efficiency and runtime to incentivize a focus on these goals. The penalty for empty predictions could also have been handled differently. Since the chosen penalty for empty predictions, i.e., false negatives, was quite harsh, it demoted the two primary affected submissions to the lower ranks. However, the investigated mitigation of reusing the last non-empty prediction could significantly increase the occurrence of false positives, which were not evaluated in the current challenge setup, where the target was assumed to be continuously visible. For this purpose, a secondary task of determining if the target is visible in the image could be implemented. The difficulty of tracking targets with motion perpendicular to the imaging plane was also identified as failure mode of even the top 5 models, resulting in subpar performance for one of the cases where such motion was present.

The final set of limitations stemmed from technical constraints of the Grand Challenge platform. While it supported smooth operation, it also introduced inefficiencies. In particular, it creates individual jobs for each case and submission. This caused significant overhead from repeatedly loading and initializing submitted algorithms. It also required processing each case as a whole sequence rather than frame by frame, unlike real-world scenarios where later frames cannot inform the prediction of earlier frames. Although submissions were required to adhere to this logic for prize eligibility, a reverse flow of information could have been required by construction. Finally, while the Grand Challenge platform ensured fair comparisons by running all submissions on identical hardware, it restricted the range of available computing options. In particular, the GPUs provided were less powerful than current state-of-the-art models. To work around this limitation, the organizers modified the real-time runtime requirements to account for the limited computational resources.

6.3. Open questions

With the increasing number of MRIgRT installations worldwide (Slotman et al., 2022; Chuong et al., 2023) and the growing clinical interest in multileaf collimator (MLC) tracking (Uijtewaal et al., 2022, 2025), research into motion tracking and adaptation is expected to intensify. A future iteration of the TrackRAD challenge could extend to other motion-tracking modalities, such as markerless X-ray tracking, to include the evaluation of tracking performance in the presence of noise, and include test data from unseen centers to better evaluate model generalization. Addressing the current limitations in dataset diversity, labeling, and computational infrastructure will be key to enabling the next generation of benchmarking efforts in MRIgRT tracking.

7. Conclusions

The TrackRAD2025 challenge successfully established a benchmark for target tracking in MRI-guided radiotherapy. The results demonstrated that finetuning promptable foundation models currently represents the state of the art for cine-MRI target tracking. Beyond identifying top-performing methods, the challenge provided valuable insights into dataset design, evaluation methodology, and technical infrastructure that will inform future developments in adaptive radiotherapy and real-time motion management.

CRediT authorship contribution statement

Tom Julius Blöcker: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Pia A.W. Görts:** Writing – review & editing, Visualization, Software, Methodology, Investigation, Conceptualization. **Yiling Wang:** Writing – review & editing, Data curation. **Elia Lombardo:** Writing – review & editing, Software, Investigation, Data curation. **Adrian Thummerer:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization. **Yu Fan:** Writing – review & editing, Data curation. **Yue Zhao:** Writing – review & editing, Data curation. **Christianna Iris Papadopoulou:** Writing – review & editing, Conceptualization. **Coen Hurkmans:** Writing – review & editing, Data curation. **Rob H.N. Tijssen:** Writing – review & editing, Data curation. **Davide Cusumano:** Writing – review & editing. **Martijn PW Intven:** Writing – review & editing, Data curation. **Pim Borman:** Writing – review & editing, Data curation. **Marco Riboldi:** Writing – review & editing, Visualization. **Denis Dudáš:** Writing – review & editing. **Hilary L. Byrne:** Writing – review & editing, Data curation. **Lorenzo Placidi:** Writing – review & editing, Conceptualization. **Marco Fusella:** Writing – review & editing, Conceptualization. **Michael Jameson:** Writing – review & editing, Data curation. **Miguel A. Palacios:** Writing – review & editing, Data curation. **Paul Cobussen:** Writing – review & editing, Data curation. **Tobias Finazzi:** Writing – review & editing, Data curation. **Shyama U. Tetar:** Writing – review & editing, Data curation. **Cornelis J.A. Haasbeek:** Writing – review & editing, Data curation. **Paul Keall:** Writing – review & editing, Conceptualization. **Matteo Maspero:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Christopher Kurz:** Writing – review & editing, Writing – original draft, Visualization, Data curation, Conceptualization. **Amparo Soeli Betancourt Tarifa:** Writing – review & editing, Software, Methodology. **Kailin He:** Writing – review & editing, Software, Methodology. **Shengqian Zhu:** Writing – review & editing, Software, Methodology. **Ying Song:** Writing – review & editing, Software, Methodology. **Guangjun Li:** Writing – review & editing, Software, Methodology. **Junjie Hu:** Writing – review & editing, Software, Methodology. **Felix Knispel:** Writing – review & editing, Software, Methodology. **Sergios Gatidis:** Writing – review & editing, Software, Methodology. **Hung Chu:** Writing – review & editing, Software, Methodology. **Jiapan Guo:** Writing – review & editing, Software, Methodology. **Maximilian Nielsen:** Writing – review & editing, Software, Methodology. **Thilo Sentker:** Writing – review & editing, Software, Methodology. **Valentin Boussot:** Writing – review & editing, Software, Methodology. **Cédric Hémon:** Writing – review & editing, Software, Methodology. **Jing Ni:** Writing – review & editing, Software, Methodology. **Konstantinos Georgas:** Writing – review & editing, Software, Methodology. **Theodoros P. Vagenas:** Writing – review & editing, Software, Methodology. **George K. Matsopoulos:** Writing – review & editing, Software, Methodology. **Guillaume Landry:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Funding acquisition, Data curation, Conceptualization.

Declaration of generative AI

While preparing this work, the authors utilized the tools available in Overleaf to correct typos and spelling errors, as well as to refine the grammar. In using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

Funding

This work was partly supported by Sichuan Science and Technology Program under Grant 2023YFH0079. The challenge has been funded thanks to a grant from Verein zur Förderung von Wissenschaft und Forschung (WiFoMed) an der Medizinischen Fakultät der Ludwig-Maximilians-Universität München e.V. awarded to Adrian Thummerer to support the computation costs. The awarded monetary prizes were provided independently by Elekta AB.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the BZKF Lighthouse Image-Guidance in Local Therapies for supporting the data collection for this dataset.

Appendix A. Participation rules and prize policies

The complete challenge design can be found in (Landry et al., 2025a).

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2026.104134.

Data availability

No new data was generated, except for the evaluation metrics and derived evaluations using the methods presented in this publication. The evaluation metrics for the submissions and the final leaderboard are available on the challenge website:

https://trackrad2025.grand-challenge.org/

The code for the presented analysis and the entirety of the challenge will be available at the following url:

https://github.com/LMUK-RADONC-PHYS-RES/trackrad2025.

The challenge dataset was partly made public as the TrackRAD2025 training dataset (Wang et al. (2025)) at the following url:

https://huggingface.co/datasets/LMUK-RADONC-PHYS-RES/TrackRAD2025

Further parts of the testing dataset were also made publicly available at the same location after the challenge closed in early 2026. Some cases (N = 20) will not be made publicly available due to privacy restrictions.

Before and during the challenge, only Elia Lombardo, Tom Julius Blöcker and Pia A.W. Görts accessed the full testing dataset.

References

Blöcker, T., Lombardo, E., Marschner, S.N., Belka, C., Corradini, S., Palacios, M.A., Riboldi, M., Kurz, C., Landry, G., 2024. Mrgrt real-time target localization using foundation models for contour point tracking and promptable mask refinement. Phys. Med. Biol. 70 (1), 015004. http://dx.doi.org/10.1088/1361-6560/ad9dad, Publisher: IOP Publishing.

Boussot, V., Hémon, C., Nunes, J.-C., Dowling, J., Rouzé, S., Lafond, C., Barateau, A., Dillenseger, J.-L., 2025. IMPACT: A generic semantic loss for multimodal medical image registration. http://dx.doi.org/10.48550/arXiv.2503.24121, URL: arXiv:2503.24121 [cs].

Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y., 2022. TransMorph: Transformer for unsupervised medical image registration. Med. Image Anal. 102615. http://dx.doi.org/10.1016/j.media.2022.102615, URL: https://www.sciencedirect.com/science/article/pii/S1361841522002432.

Chen, R., Sun, G., Li, Y., Qin, J., Benini, L., 2025. HiM2SAM: Enhancing SAM2 with hierarchical motion estimation and memory optimization towards long-term tracking. http://dx.doi.org/10.48550/arXiv.2507.07603, URL: arXiv:2507.07603 [cs].

Chuong, M.D., Clark, M.A., Henke, L.E., Kishan, A.U., Portelance, L., Parikh, P.J., Bassetti, M.F., Nagar, H., Rosenberg, S.A., Mehta, M.P., Refaat, T., Rineer, J.M., Smith, A., Seung, S., Zaki, B.I., Fuss, M., Mak, R.H., 2023. Patterns of utilization and clinical adoption of 0.35 Tesla MR-guided radiation therapy in the United States – Understanding the transition to adaptive, ultra-hypofractionated treatments. Clin. Transl. Radiat. Oncol. 38, 161–168. http://dx.doi.org/10.1016/j.ctro.2022.11.013, Publisher: Elsevier. URL: https://www.ctro.science/article/S2405-6308(22)00112-4/fulltext.

Dunn, O.J., 1964. Multiple comparisons using rank sums. Technometrics 6 (3), 241–252. http://dx.doi.org/10.1080/00401706.1964.10490181.

Green, O.L., Rankine, L.J., Cai, B., Curcuru, A., Kashani, R., Rodriguez, V., Li, H.H., Parikh, P.J., Robinson, C.G., Olsen, J.R., Mutic, S., Goddu, S.M., Santanam, L., 2018. First clinical implementation of real-time, real anatomy tracking and radiation beam control. Med. Phys. 45 (8), 3728–3740. http://dx.doi.org/10.1002/mp.13002, _eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13002, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.13002.

Huijben, E.M.C., Terpstra, M.L., Galapon, A.J., Pai, S., Thummerer, A., Koopmans, P., Afonso, M., van Eijnatten, M., Gurney-Champion, O., Chen, Z., Zhang, Y., Zheng, K., Li, C., Pang, H., Ye, C., Wang, R., Song, T., Fan, F., Qiu, J., Huang, Y., Ha, J., Sung Park, J., Alain-Beaudoin, A., Bériault, S., Yu, P., Guo, H., Huang, Z., Li, G., Zhang, X., Fan, Y., Liu, H., Xin, B., Nicolson, A., Zhong, L., Deng, Z., Müller-Franzes, G., Khader, F., Li, X., Zhang, Y., Hémon, C., Boussot, V., Zhang, Z., Wang, L., Bai, L., Wang, S., Mus, D., Kooiman, B., Sargeant, C.A.H., Henderson, E.G.A., Kondo, S., Kasai, S., Karimzadeh, R., Ibragimov, B., Helfer, T., Dafflon, J., Chen, Z., Wang, E., Perko, Z., Maspero, M., 2024. Generating synthetic computed tomography for radiotherapy: SynthRAD2023 challenge report. Med. Image Anal. 97, 103276. http://dx.doi.org/10.1016/j.media.2024.103276, URL: https://www.sciencedirect.com/science/article/pii/S1361841524002019.

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211. http://dx.doi.org/10.1038/s41592-020-01008-z, Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/s41592-020-01008-z.

Jassar, H., Tai, A., Chen, X., Keiper, T.D., Paulson, E., Lathuilière, F., Bériault, S., Hébert, F., Savard, L., Cooper, D.T., Cloake, S., Li, X.A., 2023. Real-time motion monitoring using orthogonal cine MRI during MR-guided adaptive radiation therapy for abdominal tumors on 1.5T MR-linac. Med. Phys. 50 (5), 3103–3116. http://dx.doi.org/10.1002/mp.16342, _eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.16342, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.16342.

Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C., 2023. CoTracker: It is better to track together. URL: https://arxiv.org/abs/2307.07635, arXiv:2307.07635.

Keall, P.J., Glide-Hurst, C.K., Cao, M., Lee, P., Murray, B., Raaymakers, B.W., Tree, A., van der Heide, U.A., 2022. ICRU REPORT 97: MRI-guided radiation therapy using MRI-linear accelerators. J. ICRU 22 (1), 1–100. http://dx.doi.org/10.1177/14736691221141950.

Keiper, T.D., Tai, A., Chen, X., Paulson, E., Lathuilière, F., Bériault, S., Hébert, F., Cooper, D.T., Lachaine, M., Li, X.A., 2020. Feasibility of real-time motion tracking using cine MRI during MR-guided radiation therapy for abdominal targets. Med. Phys. 47 (8), 3554–3566. http://dx.doi.org/10.1002/mp.14230, _eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.14230, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.14230.

Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30 (1/2), 81–93. http://dx.doi.org/10.2307/2332226, Publisher: [Oxford University Press, Biometrika Trust], URL: https://www.jstor.org/stable/2332226.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Amer. Statist. Assoc. 47 (260), 583–621. http://dx.doi.org/10.1080/01621459.1952.10483441, Publisher: Informa UK Limited.

Landry, G., Maspero, M., Placidi, L., Fusella, M., Cusumano, D., Hurkmans, C., Keall, P., Lombardo, E., Wang, Y., Borman, P., Jameson, M., Byrne, H., Tijssen, R., Palacios, M., Kurz, C., Riboldi, M., Dudas, D., Thummerer, A., Blöcker, T., 2025a.

Trackrad2025: Real-time tumor tracking for MRI-guided radiotherapy. http://dx.doi.org/10.5281/zenodo.15044966, Publisher: Zenodo, URL: https://zenodo.org/records/15044966.

Landry, G., Thummerer, A., Kurz, C., Hurkmans, C., Cusumano, D., Dudaš, D., Lombardo, E., Byrne, H., Placidi, L., Fusella, M., Riboldi, M., Maspero, M., Jameson, M., Palacios, M.A., Keall, P., Borman, P., Tijssen, R., Blöcker, T., Wang, Y., 2025b. Real-time tracking: The trackrad2025 challenge. In: Proceedings of the ESTRO 2025 Congress, Abstract Book. European Society for Radiotherapy and Oncology (ESTRO), Vienna, Austria, Abstract 4768, URL: https://user-swndwmf.cld.bz/ESTRO-2025-Abstract-Book/83/.

Lombardo, E., Dhont, J., Page, D., Garibaldi, C., Künzel, L.A., Hurkmans, C., Tijssen, R.H., Paganelli, C., Liu, P.Z., Keall, P.J., Riboldi, M., Kurz, C., Landry, G., Cusumano, D., Fusella, M., Placidi, L., 2024. Real-time motion management in MRI-guided radiotherapy: Current status and AI-enabled prospects. Radiother. Oncol. 190, 109970. http://dx.doi.org/10.1016/j.radonc.2023.109970, URL: https://www.sciencedirect.com/science/article/pii/S0167814023898644.

Lombardo, E., Velezmoro, L., Marschner, S.N., Rabe, M., Tejero, C., Papadopoulou, C.I., Sui, Z., Reiner, M., Corradini, S., Belka, C., Kurz, C., Riboldi, M., Landry, G., 2024b. Patient-specific deep learning tracking framework for real-time 2D target localization in MRI-guided radiotherapy. Int. J. Radiat. Oncology*Biology*Physics http://dx.doi.org/10.1016/j.ijrobp.2024.10.021.

Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B., 2025. MedSAM2: Segment anything in 3D medical images and videos. http://dx.doi.org/10.48550/arXiv.2504.03600, URL: arXiv:2504.03600 [eess].

Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 18 (1), 50–60. http://dx.doi.org/10.1214/aoms/1177730491, Publisher: Institute of Mathematical Statistics, URL: https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-18/issue-1/On-a-Test-of-Whether-one-of-Two-Random-Variables/10.1214/aoms/1177730491.full.

Mazur, T.R., Fischer-Valuck, B.W., Wang, Y., Yang, D., Mutic, S., Li, H.H., 2016. SIFT-based dense pixel tracking on 0.35 T cine-MR images acquired during image-guided radiation therapy with application to gating optimization. Med. Phys. 43 (1), 279–293. http://dx.doi.org/10.1118/1.4938096, _eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1118/1.4938096, URL: https://onlinelibrary.wiley.com/doi/abs/10.1118/1.4938096.

Palacios, M.A., Gerganov, G., Cobussen, P., Tetar, S.U., Finazzi, T., Slotman, B.J., Senan, S., Haasbeek, C.J., Kawrakow, I., 2023. Accuracy of deformable image registration-based intra-fraction motion management in magnetic resonance-guided radiotherapy. Phys. Imaging Radiat. Oncol. 26, 100437. http://dx.doi.org/10.1016/j.phro.2023.100437.

Peng, J., Stowe, H.B., Samson, P.P., Robinson, C.G., Yang, C., Hu, W., Zhang, Z., Kim, T., Hugo, G.D., Mazur, T.R., Cai, B., 2024. Inter-fractional portability of deep learning models for lung target tracking on cine imaging acquired in MRI-guided radiotherapy. Phys. Eng. Sci. Med. 47 (2), 769–777. http://dx.doi.org/10.1007/s13246-023-01371-z.

Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A., 2016. A benchmark dataset and evaluation methodology for video object segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 724–732. http://dx.doi.org/10.1109/CVPR.2016.85, ISSN: 1063-6919, URL: https://ieeexplore.ieee.org/document/7780454.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., Feichtenhofer, C., 2024. SAM 2: Segment anything in images and videos. ArXiv Preprint.

Slotman, B.J., Clark, M.A., Özyar, E., Kim, M., Itami, J., Tallet, A., Debus, J., Pfeffer, R., Gentile, P., Hama, Y., Andratschke, N., Riou, O., Camilleri, P., Belka, C., Quivrin, M., Kim, B., Pedersen, A., van Overeem Felter, M., Kim, Y.I., Kim, J.H., Fuss, M., Valentini, V., 2022. Clinical adoption patterns of 0.35 Tesla MR-guided radiation therapy in Europe and Asia. Radiat. Oncol. (London, England) 17, 146. http://dx.doi.org/10.1186/s13014-022-02114-2, URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9396857/.

Spearman, C., 1904. The proof and measurement of association between two things. Am. J. Psychol. 15 (1), 72. http://dx.doi.org/10.2307/1412159, URL: https://www.jstor.org/stable/1412159?origin=crossref.

Uijtewaal, P., Borman, P.T.S., Woodhead, P.L., Boer, H.C.J.d., Raaymakers, B.W., Fast, M.F., 2025. First experimental demonstration of magnetic resonance-guided multileaf collimator tracking for (ultra-)hypofractionated prostate radiotherapy. Phys. Imaging Radiat. Oncol. 36, http://dx.doi.org/10.1016/j.phro.2025.100828, Publisher: Elsevier, URL: https://www.phiro.science/article/S2405-6316(25)00133-2/fulltext.

Uijtewaal, P., Borman, P.T.S., Woodhead, P.L., Kontaxis, C., Hackett, S.L., Verhoeff, J., Raaymakers, B.W., Fast, M.F., 2022. First experimental demonstration of VMAT combined with MLC tracking for single and multi fraction lung SBRT on an MR-linac. Radiother. Oncol. 174, 149–157. http://dx.doi.org/10.1016/j.radonc.2022.07.004, Publisher: Elsevier, URL: https://www.thegreenjournal.com/article/S0167-8140(22)04188-3/fulltext.

Wang, Y., Lombardo, E., Thummerer, A., Blöcker, T., Fan, Y., Zhao, Y., Papadopoulou, C.I., Hurkmans, C., Tijssen, R.H.N., Görts, P.A.W., Tetar, S.U., Cusumano, D., Intven, M.P., Borman, P., Riboldi, M., Dudáš, D., Byrne, H., Placidi, L., Fusella, M., Jameson, M., Palacios, M., Cobussen, P., Finazzi, T., Haasbeek, C.J.A., Keall, P., Kurz, C., Landry, G., Maspero, M., 2025. Trackrad2025 challenge dataset: real-time tumor tracking for MRI-guided radiotherapy. Med. Phys. 52 (7), e17964. http://dx.doi.org/10.1002/mp.17964, _eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.17964, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.17964.

Warfield, S., Zou, K., Wells, W., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921. http://dx.doi.org/10.1109/TMI.2004.828354, URL: https://ieeexplore.ieee.org/document/1309714.

Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Saiz, L.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. Sci. Rep. 11 (1), 2369. http://dx.doi.org/10.1038/s41598-021-82017-6, Publisher: Nature Publishing Group, URL: https://www.nature.com/articles/s41598-021-82017-6.