

Fine-tuning Segment Anything for Real-Time Tumor Tracking in Cine-MRI

Valentin Boussot¹[0009–0003–2465–5458], Cédric Hémon¹[0009–0003–6669–5108],
Jean-Claude Nunes¹, and Jean-Louis Dillenseger¹

Univ Rennes 1, CLCC Eugène Marquis, INSERM, LTSI - UMR 1099, F-35000
Rennes, France

<https://ltsi.univ-rennes.fr/>

Abstract. In this work, we address the TrackRAD2025 challenge of real-time tumor tracking in cine-MRI sequences of the thoracic and abdominal regions under strong data scarcity constraints. Two complementary strategies were explored: (i) unsupervised registration with the IMPACT similarity metric and (ii) foundation model-based segmentation leveraging SAM 2.1 and its recent variants through prompt-based interaction. Due to the one-second runtime constraint, the SAM-based method was ultimately selected. The final configuration used SAM2.1b+ with mask-based prompts from the first annotated slice, fine-tuned solely on the small labeled subset from TrackRAD2025. Training was configured to minimize overfitting, using 1024×1024 patches (batch size 1), standard augmentations, and a balanced Dice + IoU loss. A low uniform learning rate (1×10^{-5}) was applied to all modules (prompt encoder, decoder, Hiera backbone) to preserve generalization while adapting to annotator-specific styles. Training lasted 300 epochs (12h on RTX A6000, 48GB). The same inference strategy was consistently applied across all anatomical sites and MRI field strengths. Test-time augmentation was considered but ultimately discarded due to negligible performance gains. The final model was selected based on the highest Dice Similarity Coefficient achieved on the validation set after fine-tuning. On the hidden test set, the model reached a Dice score of 0.8794, ranking 6th overall in the TrackRAD2025 challenge. These results highlight the strong potential of foundation models for accurate and real-time tumor tracking in MRI-guided radiotherapy.

1 Introduction

Medical image segmentation and registration play a crucial role in numerous clinical workflows, enabling accurate delineation of anatomical structures and pathological regions for diagnosis, treatment planning, and follow-up [3,8]. In the context of the TrackRAD2025 challenge [9], the task focuses on real-time tumor tracking in 2D cine-MRI sequences of the thoracic and abdominal regions. Only the first frame of each sequence is annotated, and the objective is to automatically segment the tumor in all subsequent frames under fast respiratory motion and strict runtime constraints (one second per sequence). This setting

reflects the clinical requirement of online motion monitoring during MRI-guided radiotherapy.

A key constraint in this challenge is the scarcity of annotated training data, which limits the applicability of conventional supervised deep learning methods that typically rely on large-scale, fully annotated datasets to reach state-of-the-art performance [5].

Given these constraints, we adopted a dual methodological strategy developed using KonfAI [1]:

- An unsupervised registration approach based on IMPACT [2], which leverages deformable image registration to propagate the initial mask across subsequent frames in the temporal sequences without requiring additional manual annotations.
- The use of foundation segmentation models, such as SAM [4] (Segment Anything Model) and MedSAM [7], which are trained on large and diverse datasets and can generalize to novel imaging modalities and anatomical sites.

The unsupervised nature of the registration approach eliminates the need for model training and parameter tuning on the challenge dataset, making it inherently robust to data scarcity. In parallel, the use of foundation models [5] provides a strong prior for anatomical segmentation [6], with the possibility of refining their performance through fine-tuning on the small annotated subset available.

This dual approach enables a direct comparison between the spatial consistency achieved through registration and the flexibility of prompted segmentation models, supporting accurate and generalizable tumor tracking in cine-MRI under limited data conditions.

2 Methods

2.1 Data

We used only the manually annotated subset of the TrackRAD2025 dataset [9]. The dataset is a multi-center collection of sagittal 2D cine-MRI sequences acquired during MRI-guided radiotherapy on two MRI-linac systems (0.35 T ViewRay MRIdian and 1.5 T Elekta Unity) across six centers. It covers thoracic (32%), abdominal (41%), and pelvic (27%) tumors. Each sequence consists of 20 to over 20000 temporal frames, sampled at 1–8 Hz depending on the vendor and acquisition protocol, capturing one or several respiratory cycles. Among the full dataset of 585 patients, 108 cases are fully annotated. In this work, we relied exclusively on the 50 publicly released annotated training cases. No unlabeled or external data were used.

2.2 Model

Unsupervised Registration with IMPACT

IMPACT is a similarity metric specifically designed for multimodal medical image registration. It leverages high-level anatomical features extracted from large,

pretrained segmentation models to guide the registration process. By integrating these semantic representations, IMPACT mitigates the limitations of traditional intensity-based similarity measures, which are often sensitive to noise, artifacts, and modality-specific intensity variations.

IMPACT was integrated into the Elastix registration framework, enabling deformable image registration between frames in a cine-MRI sequence. This approach aimed to propagate the initial mask provided for the first slice to all subsequent frames via spatial transformations estimated through registration.

While theoretically well-suited to the problem, the method faced practical constraints: the challenge imposes a one-second prediction time limit per case, which required a drastic reduction in the number of optimization iterations. Despite parameter tuning, this time constraint significantly degraded the achievable registration accuracy, preventing the full exploitation of IMPACT’s potential. Accordingly, we focus on the second approach, which aligns better with the challenge’s runtime constraints.

Tracking with SAM

For the use of segmentation foundation models, we tested many of these models that have been validated on the available annotated dataset. The SAM, MedSAM2 models and other variants were tested on the validation dataset. We then tested several prompting techniques (point sample, box, mask of the first slice) and kept the last method using the first slice. The use of test-time augmentation was also tested but did not yield any significant improvement. The same method was used for all anatomical locations and different MRI field strengths. The average inference time per cine MRI frame was approximately 10 milliseconds on our hardware.

The model SAM2.1 b+ was then fine-tuned as explained in the following section.

2.3 Training

During initial experiments using SAM for inference, we observed that the predictions were generally accurate. However, manual annotations consistently overestimated the segmented tumor volumes compared to SAM outputs. To account for this annotator bias, we fine-tuned the model on the labeled subset, enabling it to learn this discrepancy and align its predictions with the provided ground truth masks.

Fine-tuning was performed following the specific training requirements of the SAM architecture. Input patches of size 1024×1024 were extracted from the cine MRI sequences. Standard data augmentation techniques were applied, including RandomHorizontalFlip, RandomAffine, ColorJitter, and RandomGrayscale, to improve generalization and robustness. The model was trained with a batch size of 1 and supervised using a composite loss combining Dice loss (weight=1) and IoU loss (weight=1). Optimization was performed using the AdamW optimizer. Sequences of 8 consecutive frames (number of frames = 8) were provided as input.

The training procedure employed binary masks as prompts, with a learning rate of 0.0005 for the prompt network, the decoder, and the Hiera-based encoder. This relatively low learning rate was selected to preserve the pretrained capabilities of SAM while enabling it to adapt to the mask-prompt bias present in the training data.

The model was trained for 300 epochs, and the network from the final epoch was selected as the deployed model for subsequent evaluations. Training required approximately 12 hours in total on an NVIDIA RTX A6000 GPU (48 GB).

2.4 Evaluation

To assess the quality of tumor tracking in the TrackRAD2025 challenge, we employed the full set of evaluation metrics defined by the challenge organizers. These metrics aim to quantify both segmentation fidelity and the geometric accuracy of temporal tracking. They include:

- **Dice Similarity Coefficient (DSC)**: Measures the spatial overlap between the predicted and reference masks, defined as twice the intersection divided by the sum of their volumes:

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

A higher DSC (maximum value of 1) indicates stronger agreement. This metric is widely used in medical image segmentation.

- **Hausdorff Distance (HD)** and **Average Surface Distance (ASD)**: These metrics evaluate the distance between the surfaces of the predicted and reference masks. HD measures the largest surface-to-surface distance, while ASD reports the average distance, providing a more stable indicator of contour accuracy.
- **Center of Mass Distance (CD)**: The Euclidean distance between the centroids of the predicted and reference segmentations. This captures the spatial consistency of tracking over time.
- **Inference Speed**: Given that real-time performance is a critical requirement in this challenge, the inference speed of the method is recorded. The competition imposes a maximum prediction time per case (approximately one second), making this a decisive performance criterion.
- **Estimated Radiation Dose Delivery Accuracy**: Although less common in traditional segmentation benchmarks, this metric quantifies the clinical impact of segmentation on the ability to deliver the prescribed radiotherapy dose accurately. It is included in the challenge to reflect the ultimate clinical goal of real-time treatment adaptation.

This combination of classical segmentation metrics (DSC, HD, ASD), temporal tracking measures (CD, inference speed), and clinical impact assessment (dose delivery accuracy) provides a comprehensive framework for evaluating real-time tumor tracking methods.

3 Results

We evaluated the proposed method on the TrackRAD2025 validation set using the metrics described in Section 2.4, excluding dose-related metrics.

To assess the potential of registration-based tumor tracking, we evaluated our IMPACT semantic similarity metric within the Elastix framework. The registration is performed between the first frame and the current frame for which the mask is propagated. A two-resolution B-spline transformation is used, with a control-point spacing of 12 mm. At each iteration, 500 voxels are randomly sampled to evaluate the similarity metric. The optimization runs for 200 iterations per resolution level. IMPACT is configured with the SAM-2.1-Small model, using two feature layers in Jacobian mode. As shown in Table 1, IMPACT noticeably improves mask propagation accuracy compared to a baseline without registration, achieving a DSC of 0.88 and reducing the Hausdorff distance by almost 50%. However, the optimization runtime currently exceeds the real-time requirement of the TrackRAD2025 challenge (one second per cine-MRI sequence), making this approach incompatible with the official constraints. For this reason, we further focused on the fine-tuned SAM-based segmentation strategy for our final submission.

Table 1. Registration-based tumor tracking performance with unrestricted runtime (not compatible with the 1-second real-time constraint).

Method	DSC $\uparrow$	HD [mm] $\downarrow$	ASD [mm] $\downarrow$	CD [mm] $\downarrow$
Baseline	0.770	8.27	3.88	6.42
IMPACT	0.8821	4.4314	1.7868	2.3126

We then compared different SAM-based configurations and prompt strategies to identify the optimal setup for the challenge constraints. Table 2 reports the performance of different foundation model configurations. The best-performing baseline was obtained with the SAM 2.1 l configuration, which achieved a DSC of 0.916, a Hausdorff Distance of 3.26 mm, an Average Surface Distance of 1.20 mm, and a Center of Mass Distance of 1.34 mm. MedSAM-based variants performed worse across all metrics, particularly in terms of distance measures. Although pretrained SAM models already yielded strong segmentation results, a systematic overestimation of tumor volume was observed in the manual annotations compared to the model predictions. This discrepancy motivated the use of fine-tuning to adapt the model to the annotator style and improve alignment with the provided ground truth. We also evaluated an ensemble prompt strategy on the SAM2.1b+ model, combining dilated and eroded versions of the initial mask to enhance robustness. However, this approach did not outperform the best model and slightly degraded performance in some metrics.

Table 2. Comparison of different prompted foundation models for segmentation on the validation set. Best values per metric are in bold.

Model	DSC $\uparrow$	HD [mm] $\downarrow$	ASD [mm] $\downarrow$	CD [mm] $\downarrow$
Baseline	0.770	8.27	3.88	6.42
SAM 2.1 t	0.911	3.45	1.27	1.51
SAM 2.1 b+	0.915	3.41	1.29	1.46
SAM 2.1 l	0.916	3.26	1.20	1.34
MedSAM2	0.869	19.76	17.01	17.28
MedSAM2 liver	0.842	29.62	27.32	27.51
Ensemble prompt	0.908	3.51	1.29	1.39

We then investigated the effect of fine-tuning SAM2.1l to better match the annotator style. As shown in Table 3, the evaluation on the training subset showed consistent improvements across all metrics, with DSC increasing from 0.916 to 0.929 and HD decreasing from 3.26mm to 2.78mm. These gains, while expected on the training set, indicate that the model successfully adapted to the annotation style without signs of overfitting.

Table 3. Performance before and after fine-tuning on the labeled subset. Best values per metric are in bold.

Model	DSC $\uparrow$	HD [mm] $\downarrow$	ASD [mm] $\downarrow$	CD [mm] $\downarrow$
SAM 2.1 b+	0.915	3.41	1.29	1.46
Fine-tuned SAM 2.1 b+	0.929	2.78	1.00	1.16

3.1 Preliminary test set evaluation

To assess the generalization ability of the fine-tuned model, we further evaluated both versions (with and without fine-tuning) on the preliminary unseen test set provided by the organizers. As shown in Table 4, fine-tuning consistently improved all metrics, including DSC, center of mass accuracy, and shape consistency, confirming that the model adapted well to the annotator bias without overfitting.

Table 4. Preliminary test set evaluation. Fine-tuning significantly improved performance across all metrics.

Rank	Model	DSC $\uparrow$	CD [mm] $\downarrow$	Rel. D98 $\uparrow$	SD95 [mm] $\downarrow$	SDavg [mm] $\downarrow$
3rd	SAM 2.1 b+ (fine-tuned)	0.9152	1.4912	0.9803	4.2141	1.5008
5th	SAM 2.1 b+ (no fine-tune)	0.9039	1.8864	0.9312	4.9406	1.7000

3.2 Official Challenge Results

Table 5. Final performance on the hidden TrackRAD2025 test set. Rankings per metric are in parentheses.

Rank	Team	DSC $\uparrow$	CD [mm] $\downarrow$	Rel. D98 $\uparrow$	SD95 [mm] $\downarrow$	SDavg [mm] $\downarrow$
1 st	Track’n’Treat	0.8909 (1)	1.4733 (1)	0.9361 (4)	4.2171 (1)	1.6642 (1)
2 nd	Hkini_Team	0.8860 (2)	1.4858 (2)	0.9628 (1)	4.4969 (2)	1.8565 (2)
3 rd	felixknispelrwh	0.8764 (5)	1.6576 (3)	0.9250 (7)	4.5409 (3)	1.9174 (3)
4 th	CIT	0.8712 (6)	1.9564 (4)	0.9392 (2)	5.2583 (4)	2.1903 (4)
5 th	Reg’n’Track	0.8801 (3)	2.1031 (5)	0.9321 (5)	5.6407 (7)	2.2416 (5)
6 th	BreizhTrack (ours)	0.8794 (4)	2.3192 (7)	0.9363 (3)	5.6298 (6)	2.6091 (8)
14 th	IWM	0.8090 (14)	5.2809 (12)	0.7015 (14)	10.5484 (12)	4.8792 (12)

Our final submission ranked 6th overall on the TrackRAD2025 leaderboard (Table 5). The top teams exhibited very similar performance, with differences of only 0.01–0.02 in Dice and small variations in CD and SD metrics, illustrating the highly competitive nature of the challenge.

4 Discussion

Our approach ranked among the best-performing submissions in the TrackRAD2025 challenge, showing that foundation models can be effectively adapted to real-time tumor tracking in cine-MRI despite limited annotated data. The narrow performance gap between the top teams highlights the competitiveness of this benchmark.

Dice scores above 0.91 on both the training and preliminary test sets indicated good generalization during development. On the final hidden test set, the Dice dropped to 0.8794, revealing a distribution shift between the development data and the evaluation cohort. Ranking differences across both stages suggest that performance on a small publicly labeled subset does not fully anticipate robustness to variations in tumor appearance, motion patterns, and acquisition protocols.

In addition to segmentation-based tracking, we evaluated a registration strategy using the IMPACT metric. By leveraging high-level anatomical features from pretrained segmentation networks, IMPACT effectively propagated the initial mask and achieved improved performance over a no-registration baseline on the validation data. Although slightly less accurate on average, the registration proved far more reliable and clinically reassuring, maintaining anatomically consistent results even under challenging motion. In contrast, segmentation-based tracking may occasionally exhibit unpredictable failures on difficult frames, leading to local inconsistencies. However, its optimization time exceeds the one-second real-time requirement of TrackRAD2025, which precludes its use as a primary solution under the challenge constraints.

Finally, we observed that fine-tuning on a small number of annotated cases improved performance on unseen patients. However, it remains unclear whether this improvement reflects a genuine gain in segmentation accuracy or rather an adaptation of the model to specific annotation styles. Overall, these results confirm that segmentation-based strategies currently offer the best trade-off between spatial accuracy and computational efficiency for this task, particularly when enhanced through targeted fine-tuning.

5 Author contributions

Conceptualization: V. Bousso, C. Hémon. Methodology: V. Bousso, C. Hémon. Software and experiments: V. Bousso. Supervision: J.-L. Dillenseger, J.-C. Nunes. Writing – original draft: V. Bousso, C. Hémon. Writing – review and editing: all authors.

Code availability

The inference code used for our TrackRAD2025 submission is publicly available at: <https://github.com/vbousso/Trackrad2025>. We also release the fine-tuned weights of the SAM 2.1 b+ model, trained on the annotated subset of the TrackRAD2025 dataset, at: <https://huggingface.co/VBousso/Trackrad2025>. These resources allow full reproduction of our inference pipeline and challenge results.

Acknowledgment

The work presented in this article was supported by the Brittany Region through its Allocations de Recherche Doctorale framework and by the French National Research Agency as part of the VATSop project (ANR-20-CE19-0015). Additionally, it was supported by a PhD scholarship Grant from Elekta AB (C.Hémon). The authors have no relevant financial or non-financial interests to disclose. While preparing this work, the authors used ChatGPT to enhance the writing structure and refine grammar. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication’s content.

References

1. Bousso, V., Dillenseger, J.L.: Konfai: A modular and fully configurable framework for deep learning in medical imaging (2025). <https://doi.org/10.48550/ARXIV.2508.09823>
2. Bousso, V., Hémon, C., Nunes, J.C., Dowling, J., Rouzé, S., Lafond, C., Barateau, A., Dillenseger, J.L.: Impact: A generic semantic loss for multimodal medical image registration. arXiv preprint arXiv:2503.24121 (2025)

3. Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y.: Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis* **79**, 102444 (Jul 2022)
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4015–4026 (2023)
5. Kolides, A., Nawaz, A., Rathor, A., Beeman, D., Hashmi, M., Fatima, S., Berdik, D., Al-Ayyoub, M., Jararweh, Y.: Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts. *Simulation Modelling Practice and Theory* **126**, 102754 (Jul 2023). <https://doi.org/10.1016/j.simpat.2023.102754>, <https://www.sciencedirect.com/science/article/pii/S1569190X2300031X>
6. Lee, H.H., Gu, Y., Zhao, T., Xu, Y., Yang, J., Usuyama, N., Wong, C., Wei, M., Landman, B.A., Huo, Y., et al.: Foundation models for biomedical image segmentation: A survey. *arXiv preprint arXiv:2401.07654* (2024)
7. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. *Nature Communications* **15**, 654 (2024)
8. Viergever, M.A., Maintz, J.B.A., Klein, S., Murphy, K., Staring, M., Pluim, J.P.W.: A survey of medical image registration – under review. *Medical Image Analysis* **33**, 140–144 (Oct 2016)
9. Wang, Y., Lombardo, E., Thummerer, A., Blöcker, T., Fan, Y., Zhao, Y., Papadopoulou, C.I., Hurkmans, C., Tijssen, R.H., Görts, P.A., et al.: Trackrad2025 challenge dataset: real-time tumor tracking for mri-guided radiotherapy. *Medical physics* **52**(7), e17964 (2025)