# LesionLocator: Zero-Shot Universal Tumor Segmentation and Tracking in 3D Whole-Body Imaging

Supplementary Material

## Overview

This document provides supplementary details on the methods and experiments presented in the main paper:

- **Training** (Section 6): Details the architecture and parameters for both pretraining and fine-tuning of the segmentation model, followed by a discussion of the longitudinal tracking setup, including prompt propagation for multi-timepoint analysis.
- **Synthetic Data Generation** (Section 7): Outlines the lesion- and image-level augmentations used to simulate realistic disease progression and imaging variations, supporting robust training across diverse longitudinal patterns.
- **Promptable Segmentation Baselines** (Section 8): Describes the state-of-the-art promptable models benchmarked against our approach, specifically those adapted for medical imaging and segmentation.
- **Evaluation Metrics** (Section 9): Explains the metrics used for segmentation and tracking accuracy.
- **Additional Experimental Results** (Section 10): Provides further comparisons against supervised baselines on downstream tasks.

## 6. Training

### 6.1. Segmentation Model

**Pretraining.** The pretraining pipeline was implemented using the nnU-Net framework [30], specifically utilizing a ResEncL U-Net architecture [31, 33]. The model was trained for 4,000 epochs with a patch size of [192,192,192] and a batch size of 24. All images were resampled to a cubic 1mm resolution and z-score normalized. Training was performed with an initial learning rate of 1e-2, employing polynomial learning rate decay and the SGD optimizer. Following the MultiTalent strategy [100], datasets were sampled inversely proportional to the square root of the number of images per dataset, ensuring balanced training across datasets. A summary of the pretraining datasets can be found in Table 4.

**Fine-Tuning.** Fine-tuning of the promptable segmentation model was conducted using the combined lesion datasets outlined in Table 5. During data loading, images were first randomly picked, followed by random sampling of lesion instances to ensure diverse training samples. The pretrained weights were used to initialize the main body of the network, while the stem and head were randomly initialized. Fine-tuning was carried out with a reduced initial learning rate of 1e-3. Prompts were input directly at the first level of the network, concatenated with the 3D image volume. To accommodate higher-resolution images, we employed an axial spacing of 0.8mm, resulting in an overall spacing of 0.8x0.8x1mm. The patch size was accordingly adjusted to [224,224,160], and CT images were normalized following the nnU-Net protocol. The model was trained for 2,000 epochs with a batch size of 3.

### 6.2. Longitudinal Tracking

The comprehensive tracking model integrates the single timepoint segmentation network, trained in the above fashion, with the prompt propagation module. We utilize the GradICON [98] framework as a backbone for the propagation module, initializing it with pretrained weights obtained from a diverse set of image registration datasets [99]. Images from both timepoints are resampled to a uniform shape of [175, 175,175] before being fed into the prompt propagation network. We then train both prompt propagation module and segmentation model jointly on the real longitudinal data or first on the synthetic followed by fine-tuning on the real data.

Specifically, the propagation module generates a deformation field $\Phi$, which facilitates the propagation of prompts—these may include points, bounding boxes, or segmentation masks produced by the segmentation model. During training, the propagation network is provided with downsampled versions of the baseline and follow-up images, whereas the segmentation network operates on a higher-resolution cropped region centered around the propagated prompt. The center of this region of interest (ROI) is defined by a random voxel of the propagated prompt during training, while the prompt's center is used during inference. The ROI matches the segmentation network's patch size, and is extracted from the high-resolution image (0.8 x 0.8 x 1 mm) and subsequently processed by the segmentation network to generate the output mask.

We utilized PyTorch 2.3.1 and conduct experiments on NVIDIA A100 GPUs with 40GB of memory.

## 7. Synthetic Data Generation

We generate synthetic longitudinal time series data by applying instance-level lesion augmentations in combination with image-level spatial and intensity transformations to single-timepoint images.

Table 4. **Overview over all datasets used for the supervised pretraining.** This table provides a comprehensive overview of the datasets utilized for the supervised pretraining of our model. It includes a total of 47 datasets, detailing the name of each dataset, the number of images, the imaging modality employed, the specific anatomical targets, and links for direct access. These diverse datasets cover a wide range of anatomical structures and pathological conditions, ensuring a robust foundation for subsequent lesion segmentation tasks.

| Name | Images | Modality | Target | Link |
|---|---|---|---|---|
| Decatlon Task 2 [3, 92] | 20 | MRI | Heart | http://medicaldecathlon.com/ |
| Decatlon Task 3 [3, 92] | 131 | CT | Liver, L. Tumor | http://medicaldecathlon.com/ |
| Decatlon Task 4 [3, 92] | 208 | MRI | Hippocampus | http://medicaldecathlon.com/ |
| Decatlon Task 5 [3, 92] | 32 | MRI | Prostate | http://medicaldecathlon.com/ |
| Decatlon Task 6 [3, 92] | 63 | CT | Lung Lesion | http://medicaldecathlon.com/ |
| Decatlon Task 7 [3, 92] | 281 | CT | Pancreas, P. Tumor | http://medicaldecathlon.com/ |
| Decatlon Task 8 [3, 92] | 303 | CT | Hepatic Vessel, H. Tumor | http://medicaldecathlon.com/ |
| Decatlon Task 9 [3, 92] | 41 | CT | Spleen | http://medicaldecathlon.com/ |
| Decatlon Task 10 [3, 92] | 126 | CT | Colon Tumor | http://medicaldecathlon.com/ |
| ISLES2015 [63] | 28 | MRI | Stroke Lesion | http://www.isles-challenge.org/ISLES2015/ |
| BTCV [47] | 30 | CT | 13 abdominal organs | https://www.synapse.org/Synapse:syn3193805/wiki/89480 |
| LIDC [5] | 1010 | CT | Lung lesion | https://www.cancerimagingarchive.net/collection/lidc-idri/ |
| Promise12 [54] | 50 | MRI | Prostate | https://zenodo.org/records/8026660 |
| ACDC [9] | 200 | MRI | RV cavity, myocardium, LV cavity | https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html |
| ISBILesion2015 [12] | 42 | MRI | MS Lesion | https://iacl.ece.jhu.edu/index.php/MSChallenge |
| CHAOS [40] | 60 | MRI | Liver, Kidney (L&R), Spleen | https://zenodo.org/records/3431873 |
| BTCV 2 [21] | 63 | CT | 9 abdominal organs | https://zenodo.org/records/1169361#.YiDLFnXMJFE |
| StructSeg Task1 [48] | 50 | CT | 22 OAR Head & neck | https://structseg2019.grand-challenge.org |
| StructSeg Task2 [48] | 50 | CT | Nasopharynx cancer | https://structseg2019.grand-challenge.org/Home/ |
| StructSeg Task3 [48] | 50 | CT | 6 OAR Lung | https://structseg2019.grand-challenge.org/Home/ |
| StructSeg Task4 [48] | 50 | CT | Lung Cancer | https://structseg2019.grand-challenge.org/Home/ |
| SegTHOR [46] | 40 | CT | heart, aorta, trachea, esophagus | https://competitions.codalab.org/competitions/21145 |
| NIH-Pan [14, 84] | 82 | CT | Pancreas | https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT |
| VerSe2020 [52, 56, 90] | 113 | CT | 28 Vertebrae | https://github.com/anjany/verse |
| M&Ms [11, 65] | 300 | MRI | l. ventricle, r. ventricle, l. ventri. myocardium | https://www.ub.edu/mnms/ |
| ProstateX [55] | 140 | MRI | Prostate lesion | https://www.aapm.org/GrandChallenge/PROSTATEx-2/ |
| RibSeg [111] | 370 | CT | Rips | https://github.com/M3DV/RibSeg?tab=readme-ov-file |
| MSLesion [71] | 48 | MRI | MS Lesion | https://data.mendeley.com/datasets/8bctsm8jz7/1 |
| BrainMetShare [24] | 84 | MRI | Brain Metastases | https://aimi.stanford.edu/brainmetshare |
| CrossModa22 [91] | 168 | MRI | vestibular schwannoma, cochlea | https://crossmoda2022.grand-challenge.org/ |
| Atlas22 [53] | 524 | MRI | stroke lesion | https://atlas.grand-challenge.org/ |
| KiTs23 [26] | 489 | CT | Kidneys, k. Tumors, Cysts | https://kits-challenge.org/kits23/ |
| AutoPet2 [19] | 1014 | PET,CT | Lesions | https://autopet-ii.grand-challenge.org/ |
| AMOS [36] | 360 | CT,MRI | 15 abdominal organs | https://amos22.grand-challenge.org/ |
| BraTs23 [6, 7, 39, 68] | 1251 | MRI | Glioblastoma | https://www.synapse.org/Synapse:syn51156910/wiki/621282 |
| AbdomenAtlas1.0 [50, 77] | 5195 | CT | 8 abdominal organs | https://github.com/MrGiovanni/AbdomenAtlas?tab=readme-ov-file |
| TotalSegmentatorV2 [105] | 1180 | CT | 117 classes of whole body | https://github.com/wasserth/TotalSegmentator |
| Hecktor2022 [2] | 524 | PET,CT | nodal Gross Tumor Volumes (Head&Neck) | https://hecktor.grand-challenge.org/ |
| FLARE [60] | 50 | CT | 13 abdominal organs | https://flare22.grand-challenge.org/ |
| SegRap [58] | 120 | CT | 45 OARs (Head&Neck) | https://segrap2023.grand-challenge.org/ |
| SegA [37, 74, 78] | 56 | CT | Aorta | https://multicenteraorta.grand-challenge.org/data/ |
| WORD [51, 57] | 120 | CT | 16 abdominal organs | https://github.com/HiLab-git/WORD |
| AbdomenCT1K [59] | 996 | CT | Liver, Kidney, Spleen, pancreas | https://github.com/JunMa11/AbdomenCT-1K |
| DAP-ATLAS [34] | 533 | CT | 142 classes of whole body | https://github.com/alexanderjaus/AtlasDataset |
| CTORG [81] | 140 | CT | lung, brain, bones, liver, kidneys and bladder | https://www.cancerimagingarchive.net/collection/ct-org/ |
| HanSeg [75] | 42 | CT | OAR (Head&Neck) | https://han-seg2023.grand-challenge.org/ |
| TopCow [112] | 200 | CT,MRI | vessel components of CoW | https://topcow23.grand-challenge.org/ |

Table 5. **Fine-Tuning Datasets for Lesion Segmentation.** This table summarizes the 16 datasets used for fine-tuning our promptable lesion segmentation model. Each dataset contributes a collection of annotated images targeting various types of lesions. For each dataset, we provide the name, the number of images, the specific types of lesions targeted, and links to access the datasets for further exploration.

| Name | Images | Target | Link |
|---|---|---|---|
| Deep Lesion | 1093 | Various kinds of lesions | https://nihcc.app.box.com/v/DeepLesion |
| COVID-19 CT Lung | 10 | Covid -19 | https://zenodo.org/records/3757476 |
| FLARE23 Test Set | 50 | Various kinds of lesions | https://codalab.lisn.upsaclay.fr/competitions/12239 |
| KiTS | 488 | Kidney Lesions | https://kits-challenge.org/kits23/ |
| LIDC | 1010 | Lung Lesions | https://www.cancerimagingarchive.net/collection/lidc-idri/ |
| LNDb | 229 | Lymph nodes | https://lndb.grand-challenge.org |
| MSD Colon | 126 | Colon Lesions | http://medicaldecathlon.com/ |
| MSD Hepatic Vessels | 303 | Liver Lesions | http://medicaldecathlon.com/ |
| MSD Liver | 118 | Liver Lesions | http://medicaldecathlon.com/ |
| MSD Lung | 63 | Lung Lesions | http://medicaldecathlon.com/ |
| MSD Pancreas | 281 | Pancreas Lesions | http://medicaldecathlon.com/ |
| NIH Lymph | 176 | Lymph nodes | https://www.cancerimagingarchive.net/collection/ct-lymph-nodes/ |
| NSCLC Pleural effusion | 78 | Pleural effusion | https://www.cancerimagingarchive.net/analysis-result/plethora/ |
| NSCLC Radiomics | 503 | Lung Lesions | https://www.cancerimagingarchive.net/collection/nsclc-radiomics/ |
| autoPET | 500 | Melanoma | https://autopet-ii.grand-challenge.org/ |
| COVID-19-20 | 199 | Covid-19 | https://covid-segmentation.grand-challenge.org/COVID-19-20/ |

### 7.1. Lesion-Level Augmentations

To simulate random disease progression, we adapt the anatomy-informed transformation approach [43] to model lesion growth and shrinkage, creating realistic synthetic longitudinal data. Specifically, we construct deformation fields $V$ around each lesion by computing the gradient of a Gaussian kernel $G_{\sigma s}$ convolved with a lesion indicator function $S_{lesion}$, i.e. the lesion ground truth mask, scaled by an amplitude $A$:

$$V = \nabla(G_{\sigma s} * S_{lesion}(x, y, z)) \cdot A(x, y, z).$$

To introduce variability in progression, we modulate the amplitude $A$ with a location-dependent random field $r(x, y, z) \sim \mathcal{U}(r_{\text{min}}, r_{\text{max}})$, then smooth this field using a Gaussian kernel $G_{\sigma r}$:

$$A(x, y, z) = A \cdot (G_{\sigma r} * r(x, y, z)).$$

Given the significant size variability of lesions compared to surrounding anatomical structures, fixed transformation parameters can distort smaller lesions or inadequately alter larger ones. To address this, we employ a multi-stage approach, applying a sequence of moderate transformations with parameters adapted to lesion size:

- $\mathbf{G_{\sigma s}}$: Size of the Gaussian kernel used to blur the lesion segmentation, adapted based on lesion size, with values ranging from $4 - 5.5$.
- $\mathbf{A}$: Initial amplitude for lesion dilation, randomly sampled across the entire image from the set $[-22, -18, 15, 25]$. Negative values simulate lesion shrinkage, while positive values induce growth.
- $\mathbf{r(x, y, z)}$: Random voxel-wise scaling field for amplitude $A$, with values drawn from the range $(-3.5, 3.5)$. This ensures spatial variability, so that the lesion grows or shrinks non-uniformly across 3D space.
- $\mathbf{G_{\sigma r}}$: Gaussian kernel size, fixed at 3, used to smooth the random modulation field $r(x, y, z)$.

This transformation approach ensures robust, size-sensitive adjustments to lesions, creating realistic variations in lesion shape and size over time.

### 7.2. Image-Level Augmentations

We further enhance these lesion-level alterations with image-level intensity and spatial augmentations to simulate real-world examination variability. In our augmentation pipeline, we employ the `batchgeneratorsv2` [29] package to streamline the application of spatial and intensity transformations. Below, we detail each transformation included in the pipeline:

- **Spatial Transform**: We utilize elastic deformations, rotations, scaling, and translations to introduce realistic spatial variations. Key parameters include:
  - **Elastic Deformations**: Applied with a probability of 1.0 to simulate structural variability, with `elastic_deform_scale` and `elastic_deform_magnitude` set to $(0.05, 0.05)$.
  - **Rotation**: Random rotations in the range $(-5°, 5°)$ are applied with a probability of 1.0, introducing slight angular variations.
  - **Scaling**: Applied with a probability of 0.5, using scaling factors drawn from $(0.95, 1.05)$, and set to synchronize across all axes for uniform scaling.
  - **Translation**: Minor translations within the range $(-5, 5)$ pixels are applied with a probability of 1.0 to emulate slight spatial shifts in image positioning.
- **Gaussian Noise Transform**: We add Gaussian noise to simulate varying noise levels across imaging sessions. Noise variance is sampled from $(0, 0.05)$ and applied independently across channels, with a probability of 1.0.
- **Gaussian Blur Transform**: Gaussian blurring with a sigma range of $(0.1, 0.2)$ is applied with a probability of 0.1 to replicate the effects of lower scan quality or minor out-of-focus regions. This transform is applied in an unsynchronized manner across channels and axes to maintain realistic variability.
- **Multiplicative Brightness Transform**: Brightness adjustments are applied with a probability of 0.15 to emulate diverse lighting conditions, with brightness multipliers drawn from the range $(0.75, 1.25)$.
- **Contrast Transform**: Contrast is adjusted with a probability of 0.15 to simulate different imaging conditions. Contrast levels are sampled from the range $(0.75, 1.25)$ and applied while preserving the original intensity range to prevent artifacts.

We provide additional examples of synthetically generated longitudinal images in Fig. 7.

## 8. Promptable Segmentation Baselines

The **Segment Anything Model** (SAM) by META is a leading model from the natural image domain that has inspired numerous researchers to adapt it for radiological medical imaging. While it was trained on 1 billion masks and 11 million images, it did not focus explicitly on radiological data. SAM was the first to popularize interactive segmentation approaches [41].

**MedSAM** is a tailored adaptation of SAM, fine-tuned on 1,570,263 image-mask pairs specifically from the medical domain. Unlike its predecessors, MedSAM is limited to box prompts [61].

**SAM-Med2D** is a SAM ViT-b model with additional

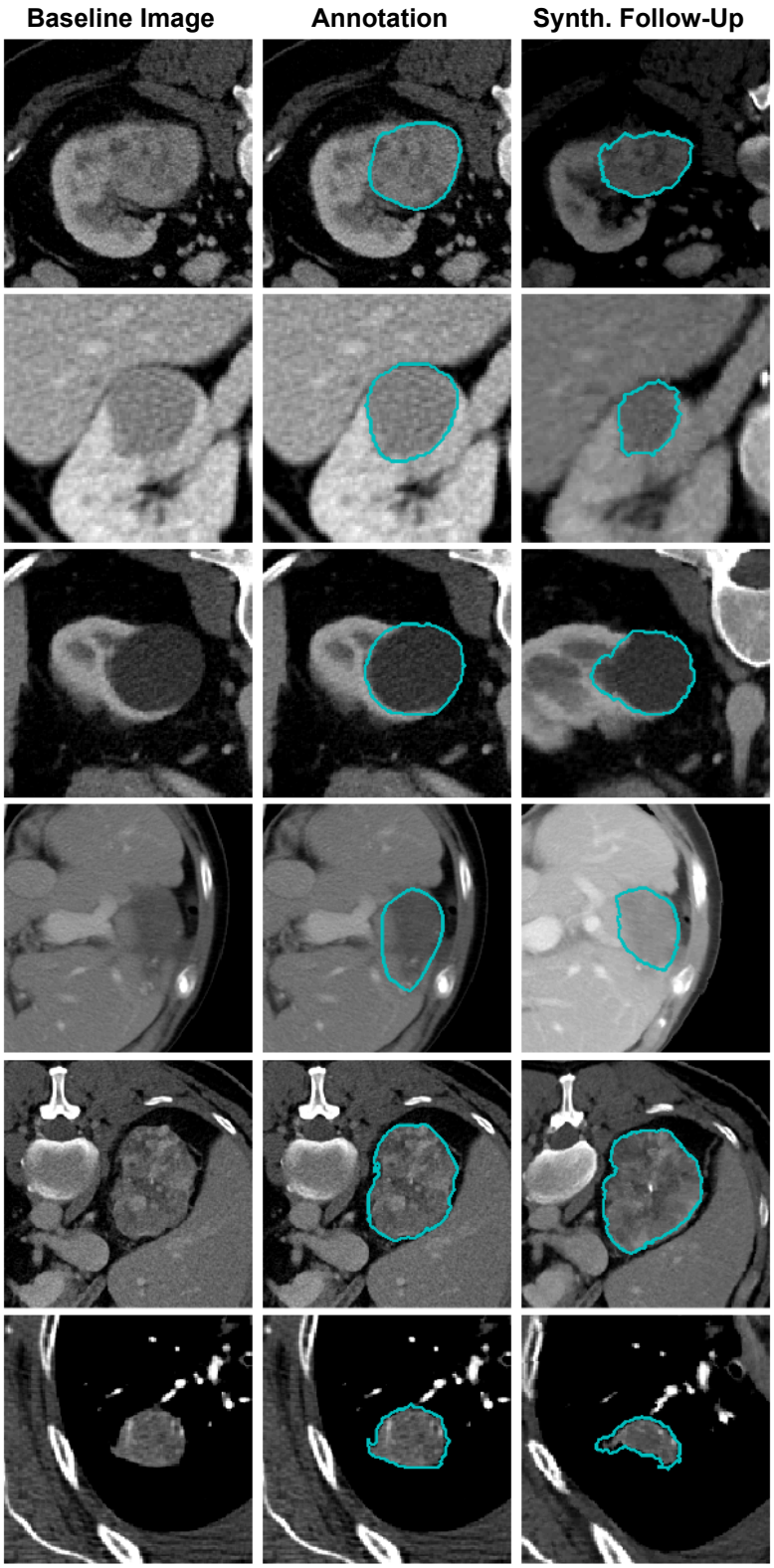


Figure 7. More examples from the synthetic dataset used to augment the training process for lesion tracking. The dataset simulates disease progression through random lesion progression, based on anatomy-informed transformations and image augmentations.

adapter layers in the image encoder. It was fine-tuned on 4.6 million images and 19.7 million masks from the medical domain using boxes and clicks.[13].

**ScribblePrompt** is a state-of-the-art medical segmentation model that supports prompting via points, boxes, or scribbles. It offers the flexibility of utilizing either a UNet or SAM model (ViT-b) backbone; we opted for the UNet backbone due to its superior performance in the performed user study. This model is trained on a comprehensive dataset of 65 diverse medical sources, encompassing a wide range of both healthy anatomical structures and various lesions [107].

**SAM2** extends SAM, enhancing its capabilities by incorporating support for video data and increasing the training dataset size [79]. In our experiments, we utilize the SAM2.1 Hiera Base Plus checkpoint and evaluate it in a 2D+t configuration, treating axial slices as individual images and interpreting the z-dimension as the temporal axis.

**SAM-Med3D** introduces a transformer-based 3D image encoder, 3D prompt encoder, and 3D mask decoder. The original model was trained from scratch on 22,000 3D images and 143,000 corresponding 3D masks. **SAM-Med 3D Turbo** is an enhanced version of SAM-Med 3D, trained on a more extensive dataset collection consisting of 44 datasets for improved performance, which we use in our comparisons. It supports both point and mask prompts [104].

**NVIDIA VISTA** is a 3D segmentation model that supports point prompts in conjunction with class prompts for 127 common human anatomical structures and various lesion types. The model leverages SegResNet [72] as its backbone CNN, enhanced by SAM's prompt encoder. It was trained on a comprehensive dataset comprising 11,454 CT volumes, both private and public, which include real and pseudo labels [25].

**SegVol** is an interactive 3D segmentation model that utilizes a 3D adaptation of the Vision Transformer (ViT) architecture. It was initially trained on 96,000 unlabelled CT images and subsequently fine-tuned using 6,000 labeled CT images. SegVol supports both point and bounding box prompts as spatial inputs, as well as corresponding text prompts that describe the class. In our experiments, we prompt with "lesion" or "tumor" which increased results [18].

**ULS model**: Developed for the Universal Lesion Segmentation challenge [15], the ULS model is specifically tailored for lesion segmentation. It was trained on 38,693 lesions derived from 3D CT scans covering the entire body. However, it does not function as a traditional promptable model, as it operates on a fixed region of interest (ROI) with the expected lesion centered for segmentation. Consequently, it can only be utilized as a point model by employing center points as inputs.

In addition to these models, other notable promptable models exist, including 3D Sam Adapter [23] and Prism [49]. However, these models operate in a closed-set manner, having been trained exclusively on specific datasets without the capability to segment arbitrary prompted classes. Therefore, they were excluded from our evaluations.

## 9. Evaluation

For evaluating segmentation and tracking performance, we employ a range of metrics that capture both accuracy and robustness in handling diverse lesion sizes and positions. The single-timepoint segmentation models are evaluated on six held-out lesion segmentation datasets, as detailed in Table 6. This dataset collection includes a multi-timepoint, in-house annotated whole-body melanoma dataset, which we use for all tracking model experiments through 5-fold cross-validation, addressing the lack of suitable public longitudinal datasets.

### 9.1. Single-Timepoint Segmentation Metrics

We evaluate segmentation performance using two primary metrics:

- **Dice Score**: Measures the overlap between the predicted and ground truth lesion masks.
- **Normalized Surface Dice (NSD)** with a 2mm tolerance: Ensures precise boundary delineation, accounting for varying lesion sizes by focusing on surface-level deviations. This metric is particularly suited for handling both small and large lesions.

### 9.2. Tracking Metrics

For tracking across longitudinal scans, we prompt the previous timepoint and evaluate the model's performance on the follow-up segmentation using the following metrics:

- **Center Point Matching (CPM@25)**: The percentage of ground truth and predicted lesion center points within a 25mm distance, reflecting the accuracy of lesion localization over time. We follow the 25mm threshold used in related work for consistency.
- **Dice@25**: The Dice score for lesions with center points correctly matched within 25mm, capturing segmentation quality for accurately tracked lesions.
- **Mean Euclidean Distance (MED)**: The average Euclidean distance between predicted and ground truth le-

Table 6. **Test Datasets.** This table summarizes the 6 datasets used for evaluating our proposed model. The datasets encompass a large set of annotated images from various types of lesions and institutions. For each dataset, we provide the name, the number of images, the specific types of lesions targeted, and links to access the datasets.

| Name | Images | Target | Link |
|---|---|---|---|
| Liver Metastases | 171 | Colorectal Cancer | www.cancerimagingarchive.net/collection/colorectal-liver-metastases/ |
| Adrenal-ACC-Ki67-Seg | 53 | Adrenocortical Carcinoma | www.cancerimagingarchive.net/collection/adrenal-acc-ki67-seg/ |
| HCC-TACE-Seg | 66 | Primary Liver Cancer | www.cancerimagingarchive.net/collection/hcc-tace-seg/ |
| Lnq2023 | 393 | Malignant Lymph Nodes | lnq2023.grand-challenge.org/ |
| RIDER Lung CT | 55 | Lung Cancer | www.cancerimagingarchive.net/collection/rider-lung-ct/ |
| Whole-body Melanoma | 159 | Metastatic Melanoma | Private |

| Dim | Model | Prompt | Colorectal Liver Tumor | Adrenal Tumor | Primary Liver Cancer | Lymph Node Metastases | Lung Tumor | Whole-body Melanoma | Avg. Dice |
|---|---|---|---|---|---|---|---|---|---|
| 3D | nnUNet [30] | - | 64.09 | 89.03 | 72.27 | 43.34 | 72.05 | 62.01 | 67.13 |
| | ***Ours*** | point | 75.38 | 89.10 | 78.39 | 75.60 | 77.18 | 82.58 | **79.71** |
| | | box | 74.05 | 92.04 | 85.71 | 80.63 | 82.51 | 84.66 | **83.26** |
| Dice Inter-Rater Variability | | | 76 [35] | - | 84 [35] | 80 [44] | 81-85 [42] | 80-85[28] | |

Table 7. **Performance Comparison Against Supervised Segmentation.** This table compares the segmentation performance of our model with nnUNet, a leading supervised medical segmentation model trained specifically on each test dataset and evaluated via 5-fold cross-validation. Remarkably, our model, despite never being trained on these held-out datasets, achieves higher Dice scores across all lesion types by leveraging either point or box prompts. Results are benchmarked against human inter-observer variability, offering an upper bound reference.

sion center points, providing a direct measure of tracking precision. Lesions without a corresponding match in the ground truth or prediction are excluded from this calculation.

- **Total Dice Score**: The overall Dice score across all tracked lesions, assessing the model's ability to maintain segmentation quality over time, including missed or wrongly matched lesions.

All tracking metrics are averaged by patient. First, the average over all lesions of a particular scan is calculated, and then weighted by the number of scans per patient to account for variability in the number of available scans per patient.

## 10. Additional Results

**Zero-Shot Segmentation Performance Exceeds Supervised Models.** We benchmarked our zero-shot promptable segmentation model against nnUNet [30], a leading supervised segmentation framework that has consistently set high standards in medical image segmentation [32]. To establish a robust comparison, nnUNet was trained independently on each of our six benchmark datasets, ensuring it had full access to the specific lesion types and image distributions within each dataset (see Tab. 7). Remarkably, despite nnUNet's access to the dataset from each specific lesion type, our zero-shot model outperformed it by over 15 Dice points on average, a significant margin that underscores the versatility and generalization capabilities of our approach.

Our model achieved these results without any prior exposure to the datasets or lesion-specific information, relying solely on prompts such as points or bounding boxes to localize regions of interest. This not only highlights the model's zero-shot proficiency but also its robustness across varied anatomical contexts, from colorectal liver tumors to whole-body melanoma. In addition, our model's performance in zero-shot settings closely approaches or even reaches inter-rater variability levels reported in literature, further reinforcing its reliability and potential as a scalable solution in clinical scenarios where labeled data may be limited or unavailable.

**Consistently High Tracking Performance Irrespective Of Prompt.** LesionLocator achieves robust and consistent temporal tracking accuracy, as demonstrated in Fig. 5 of the main paper. In this figure, the initial Dice distribution on the baseline scan shows strong performance using box prompts in the segmentation module. To complement this, Appendix Fig. 8 illustrates a similar initial distribution with less informative point prompts, which perform slightly lower overall but still demonstrate high accuracy. The bar plots for subsequent timepoints show that LesionLocator consistently achieves high Dice scores using autoregressive mask prompts, even when trained exclusively on consecutive image pairs. This highlights the generalizability of our longitudinal training approach. With minimal performance

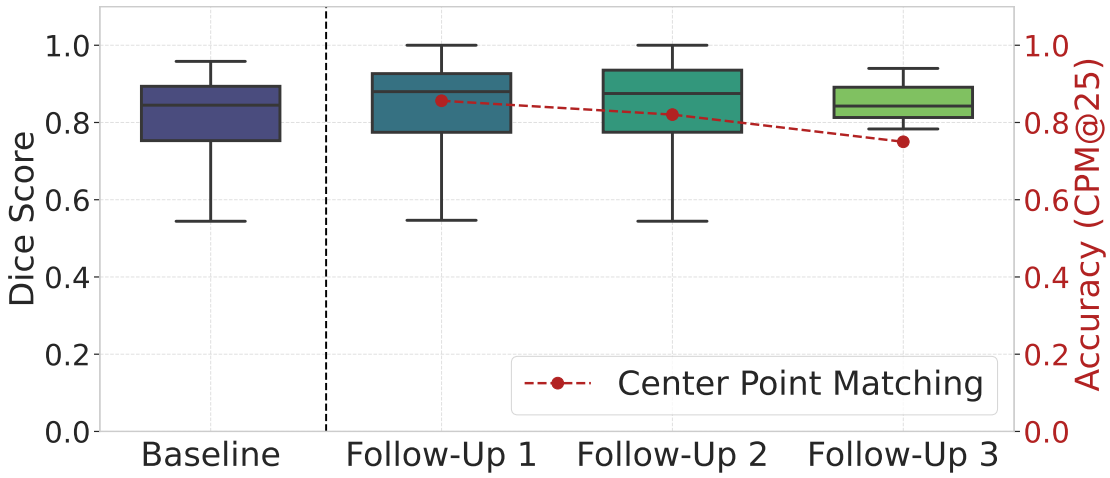


Figure 8. **Consistent Lesion Tracking Performance Over Time Using Point Prompts.** Similar to Fig. 5, we show the initial Dice score distribution for the baseline scan using LesionLocator's segmentation model with *point prompts*. For follow-up scans, tracking is performed autoregressively, as proposed, using prior masks as prompts. Tracking accuracy relative to the baseline is measured as CPM@25 (lesion matches within 25mm) with corresponding Dice@25 (Dice score of matched lesions). Similar to using box prompts in the first image, the Dice for matched lesions remains consistently high, with matching accuracy above 80% and only a slight decrease over time. Note: Only a single patient in the dataset has a Follow-Up 3 scan, so this distribution is based on one scan with 4 lesions, of which 3 were correctly matched.

| Data | Model | Dice@25↑ | MED↓ |
|---|---|---|---|
| from paper | Yan et.al [110] *(point tracker)* | - | 6.92 |
| | Registration of $t-1$ Seg | 36.10 | 8.71 |
| | Def. Register + LesLoc Seg | 68.54 | 5.79 |
| | ***Ours (LesionLocator)*** | **79.02** | **3.12** |
| unseen OOD | Yan et.al [110] *(point tracker)* | - | 9.07 |
| | Hering et.al [27](*best baseline*) | 64.42 | 6.39 |
| | ***Ours (LesionLocator)*** | **76.55** | **5.13** |

Table 8. Ablations and diffuse-lesion diverse test set results.

degradation across multiple timepoints, LesionLocator excels in lesion matching (CPM@25), ensuring reliable and sustained tracking throughout a sequence of scans irrespective of initial prompt type.

**Comparison to Registration-Based Cross-Time Segmentation.** Tab. 8 (top) compares our approach against classical Elastix [64], which warps prior segmentations to the follow-up scan using its deformation field, and a two-step approach that uses Elastix to warp the prompt instead, then feeds it into LesionLocator's segmentation module. The results clearly indicate the superiority of our combined training approach.

**Robustness.** To further evaluate out-of-distribution performance, we extended our experiments to assess robustness under real-world conditions. Specifically, our clinicians additionally annotated patients with diffuse, challenging-to-segment lesions from two centers. These cases feature varying resolutions, implant artifacts and were unseen during training. The results shown in Tab. 8 (bottom) demonstrate that our method, despite the expected decrease, maintains robust performance.