LETT-NeXt: A Lightweight RECIST-Guided Model for 3D CT Lesion Segmentation

Sebastian Aas[0009−0006−2453−1289],
Elias Stenhede[0009−0005−2654−4553], and
Arian Ranjbar[0000−0002−0422−2255]

Medical Technology & E-health, Akershus University Hospital, Lørenskog, Norway
Faculty of Medicine, University of Oslo, Oslo, Norway
sebastian.aas@ahus.no

Abstract. RECIST diameter measurements are widely used for tumor response assessment, but they provide only a limited 2D description of lesion extent. We present LETT-NeXt, a lightweight RECIST-guided model that predicts 3D lesion masks from CT volumes and RECIST markers for the CVPR 2026 Foundation Models for Pan-cancer Segmentation in CT Images competition. LETT-NeXt extracts a RECIST-centered regional crop, encodes the RECIST line and endpoints as two prompt channels, and concatenates them with the CT input. A compact MedNeXt-v2 encoder–decoder predicts the lesion mask, followed by prompt-aware component selection and adaptive AutoZoom inference. On the public validation set, LETT-NeXt achieved a Dice Similarity Coefficient (DSC) of 79.4 ± 10.1 and a Normalized Surface Dice (NSD) of 72.3 ± 16.2. On the hidden test set, it achieved a DSC of 73.9 and an NSD of 67.3, corresponding to a challenge score of 70.6%. On the public validation mirror, LETT-NeXt completed CPU inference in 6.9 ± 3.0 s per case with a peak memory use of 3.6 GB. Code is available at `https://github.com/Ahus-AIM/lett-next`.

Keywords: Computed Tomography · Interactive Tumor Segmentation

1 Introduction

RECIST v1.1 [4] is widely used to standardize tumor response assessment in clinical practice and clinical trials. Its diameter-based measurements are practical and easy to acquire, but a small number of 2D measurements cannot fully represent the 3D shape and spatial extent of a lesion. Volumetric lesion segmentation can describe tumor burden more completely, and previous studies suggest that volumetric analysis may provide clinically useful information beyond RECIST in settings such as lung cancer response assessment and breast cancer treatment response prediction [5,16]. However, full 3D lesion annotation is difficult to scale in time-constrained clinical workflows.

The CVPR 2026: Foundation Models for Pan-cancer Segmentation in CT Images competition addresses this gap by asking models to predict a 3D lesion

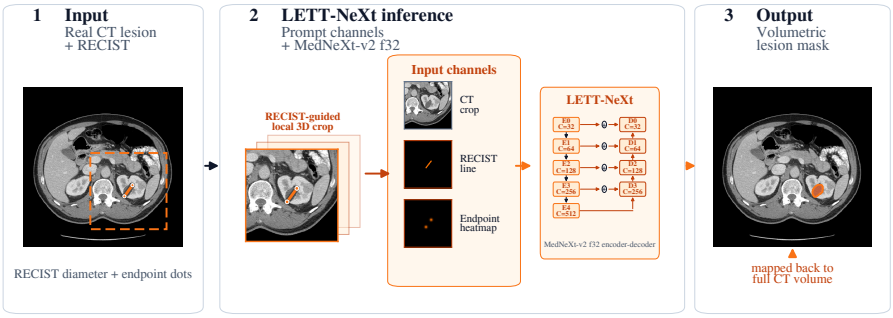

Fig. 1: LETT-NeXt converts a RECIST marker into a volumetric lesion mask.

mask from a CT volume and a RECIST marker. This setting requires accurate local segmentation from sparse spatial input, while lesions vary in size, shape, appearance, and anatomical location. The challenge also emphasizes efficient inference under strict computational constraints, making runtime and memory use important design considerations.

1.1 Related work

Medical image segmentation has long been dominated by encoder–decoder architectures, most notably U-Net [19]. 3D variants such as SegResNet [17] extend this family to volumetric segmentation, while nnU-Net [9] showed the importance of adapting preprocessing, training, inference, and postprocessing to each dataset. U-Net-style models therefore remain strong baselines for medical segmentation, especially when combined with task-specific configuration [15].

Recent ConvNeXt-style models further show that convolutional backbones remain competitive for 3D medical image segmentation. MedNeXt [21] adapts ConvNeXt [13] principles to volumetric segmentation through a fully convolutional encoder–decoder with residual upsampling and downsampling blocks. MedNeXt-v2 [20] extends this line of work by scaling 3D ConvNeXt backbones for supervised representation learning and adding volumetric global response normalization. Its evaluation against several strong public segmentation models [24,8,6,**?**,7] makes it a useful reference point for modern 3D medical segmentation backbones.

Promptable segmentation conditions prediction models on user- or task-provided spatial information. General segmentation models established this paradigm for natural images, videos, and concept-level prompts [11,18,3], while medical variants adapt promptable segmentation to medical images and volumetric data [14,23,**?**]. In 3D medical segmentation, nnInteractive is especially relevant

because it represents spatial prompts as additional input channels and introduces AutoZoom for adaptive field-of-view refinement [10].

ENSAM introduced a promptable 3D medical image segmentation framework with a SegResNet-based image encoder, prompt encoder, and mask decoder connected through latent cross-attention [22]. Lite ENSAM adapted this framework to RECIST-conditioned CT lesion segmentation by using RECIST markers as sparse prompts and by reducing the computational footprint for CPU-based inference [2]. Together, these models provide the main starting point for the present work: prompt-conditioned volumetric segmentation, RECIST-guided lesion localization, and lightweight inference.

Building on ENSAM and Lite ENSAM, we present LETT-NeXt, a lightweight RECIST-guided 3D lesion segmentation model. LETT-NeXt replaces latent prompt decoding with direct RECIST prompt-channel conditioning and uses a compact MedNeXt-v2 encoder–decoder for efficient local segmentation. During training, an auxiliary anatomy–tumor head provides additional supervision. At inference, RECIST-guided regional crops and AutoZoom maintain a local field of view, while still allowing adaptive expansion when needed.

2 Method

The method section first describes RECIST-guided regional cropping and prompt encoding, then introduces the MedNeXt-v2 segmentation model, and finally describes prompt-aware postprocessing and AutoZoom inference.

2.1 Preprocessing

All CT volumes were resampled to a spacing of $2.4 \times 1.0 \times 1.0$ mm in z, y, x order. For each RECIST annotation, we extracted a RECIST-centered regional crop of size $72 \times 160 \times 160$ voxels. Predictions produced in crop space were mapped back to the original image grid after inference.

CT intensities were clipped to $[-999, 255]$ HU and normalized by subtracting the training-set mean $\mu = 69.86$ and dividing by the training-set standard deviation $\sigma = 200.88$. The normalized crop is denoted by $\tilde{\mathbf{I}}_{\mathrm{CT}}$.

LETT-NeXt conditions the segmentation model on RECIST annotations by concatenating two prompt channels with the normalized CT crop. The first channel, $\mathbf{P}_{\mathrm{line}}$, is a binary RECIST-line mask. The two RECIST endpoints are connected on the axial RECIST slice and thickened with disks of radius $r = 2$ voxels. Voxels covered by the thickened line are assigned value 1, and all other voxels are assigned value 0.

The second channel, $\mathbf{P}_{\mathrm{endpoints}}$, is an endpoint heatmap. For voxel coordinate $\mathbf{v} = (z, y, x)$ and RECIST endpoints $\mathbf{p}_1, \mathbf{p}_2$ in crop coordinates, the heatmap is

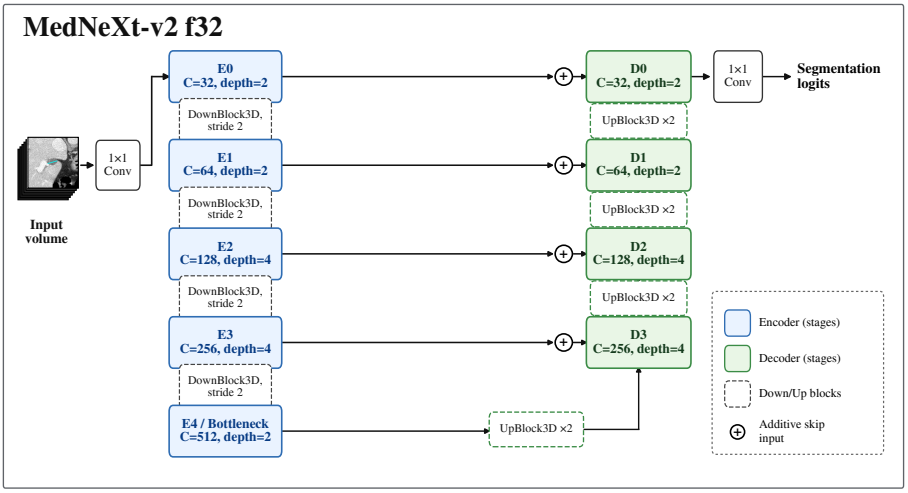

Fig. 2: Overview of the MedNeXt-v2 f32 architecture used in LETT-NeXt. The network has five encoder scales, four decoder scales, skip connections between corresponding resolution levels, and encoder channel widths of 32, 64, 128, 256, and 512.

defined as

$$\mathbf{P}_{\text{endpoints}}(\mathbf{v}) = \max_{i \in \{1,2\}} \exp\left(-\frac{\|\mathbf{v} - \mathbf{p}_i\|_2^2}{2\sigma_{\text{end}}^2}\right),$$

where $\sigma_{\text{end}} = 2.0$ voxels controls the spatial spread of the endpoint prompt.

The final network input is the three-channel tensor

$$\mathbf{X} = \left[\tilde{\mathbf{I}}_{\text{CT}}, \mathbf{P}_{\text{line}}, \mathbf{P}_{\text{endpoints}}\right].$$

2.2 Proposed Method

Figure 2 illustrates the LETT-NeXt segmentation network. LETT-NeXt uses the three-channel input $\mathbf{X}$ from Section 2.1 and predicts a binary lesion mask with a compact MedNeXt-v2 f32 encoder–decoder [20]. The network follows a 3D U-Net-like topology with skip connections between corresponding resolution levels and contains approximately 6.92 million trainable parameters. Additional block-level details are provided in Appendix B.

Auxiliary anatomy supervision During training, LETT-NeXt used a training-only auxiliary anatomy–tumor head in addition to the main RECIST-guided lesion segmentation head. This was motivated by anatomy-aware supervision in

PanTS [12], and was used to encourage anatomy-aware feature learning around the prompted lesion.

The auxiliary head was implemented as a $1 \times 1 \times 1$ convolution attached to the final decoder feature map. It predicted a local anatomy–tumor map for PanTS-supervised training crops, while the main RECIST-guided head predicted a binary lesion mask.

The auxiliary loss was computed only over voxels with valid PanTS-derived auxiliary labels. It combined multi-class cross-entropy and multi-class Dice loss over the auxiliary anatomy–tumor label space:

$$\mathcal{L}_{\text{aux}} = 0.5\,\mathcal{L}_{\text{CE}}\left(\hat{\mathbf{Y}}_{\text{aux}}, \mathbf{Y}_{\text{aux}}\right) + 0.5\,\mathcal{L}_{\text{Dice}}^{\text{multi}}\left(\hat{\mathbf{Y}}_{\text{aux}}, \mathbf{Y}_{\text{aux}}\right).$$

For crops without valid auxiliary labels, the auxiliary term was set to zero. At inference, only the main lesion head was used, and the auxiliary logits were ignored.

2.3 Postprocessing

LETT-NeXt performed inference independently for each RECIST marker in a CT scan. Each prompt corresponds to one target lesion, so the pipeline produced one lesion prediction per RECIST marker. For each prompt, the corresponding input tensor $\mathbf{X}$ was passed through the MedNeXt-v2 f32 model. The sigmoid output was thresholded at 0.35 to obtain candidate lesion voxels. This value was selected by a validation-set threshold sweep.

To ensure that each prediction corresponded to a single prompted lesion, connected components were computed on the thresholded crop-space mask. The component most consistent with the RECIST marker was retained. Selection first used the component with the largest overlap with the RECIST marker region. If no component overlapped the prompt region, the component closest to the prompt center was selected instead. The selected crop-space mask was mapped back to the original image grid and assigned the corresponding RECIST label ID in the final output mask.

AutoZoom AutoZoom is designed for cases where the initial RECIST-centered crop may not fully contain the lesion. Small crops preserve more image detail, but lesions that extend beyond the crop boundary can be truncated. Following the prompt-guided local inference strategy of nnInteractive [10], LETT-NeXt performs a second inference pass only when the first-pass prediction touches the crop boundary.

For the second pass, a larger physical crop is extracted around the same RECIST marker, resized to the fixed network input size, and passed through the same model. The expanded prediction replaces the initial prediction, after which the same thresholding, component selection, and mapping procedure is applied.

3 Experiments

3.1 Dataset and evaluation measures

The development data were provided by the CVPR 2026: Foundation Models for Pan-cancer Segmentation in CT Images competition and hosted on Hugging Face [1]. The challenge-provided training and validation partitions are referred to as the training and validation sets.

The training data contained 25,112 records in total: 24,311 primary challenge records and 801 PanTS records used only for auxiliary supervision. The primary records came from 9,968 unique cases. Training was performed at the lesion-prompt level: each competition record paired one CT volume with one target lesion mask, voxel spacing metadata, and the corresponding RECIST marker. Therefore, cases with multiple target lesions could contribute multiple training records. PanTS records were used only for auxiliary anatomy–tumor supervision.

The validation set contained 49 CT cases with 84 annotated target lesions and RECIST markers. At inference, prompt-level predictions were merged into one labeled output mask per CT case. Unless otherwise stated, validation metrics were computed per CT case and reported as mean ± SD over the 49 cases. Hidden test results were obtained from the challenge evaluation server, which reported aggregate metrics only.

Development results were measured using the challenge score, defined as the mean of Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD):

$$\text{Score} = 0.5 \cdot \text{DSC} + 0.5 \cdot \text{NSD}. \tag{1}$$

Unless otherwise stated, DSC, NSD, and challenge score are reported on a 0–100 scale, obtained by multiplying the corresponding $[0, 1]$ metric values by 100. Absolute differences are reported in percentage points.

3.2 Implementation Details

LETT-NeXt was trained using a mixed RECIST-centered crop recipe. Standard crops were centered on the RECIST marker and endpoints, matching the prompt geometry used at inference. Enlarged-field crops used the same center but sampled a larger physical field of view before resizing to the fixed input size. Bounding-box-fitted crops were used only when a standard RECIST-centered crop risked cutting off part of the lesion; in these cases, the crop was centered on the lesion bounding box and adjusted to keep the full lesion inside the patch.

The nominal sampling recipe was 60% standard RECIST marker crops, 30% enlarged-field RECIST crops, and up to 10% bounding-box-fitted crops. If bounding-box-fitted sampling was not applicable, sampling fell back to the standard RECIST marker crop. All crop types enforced inclusion of the RECIST endpoints with a 15 mm margin.

Table 1: Training protocol for LETT-NeXt.

Parameter	**Value**
Backbone	MedNeXt-v2 f32
Input crop size	$72 \times 160 \times 160$ voxels, z, y, x
Target spacing	$2.4 \times 1.0 \times 1.0$ mm, z, y, x
Training data	9,968 cases; 25,112 records
Validation set	49 CT cases; 84 RECIST markers
Epochs	25
Optimizer	AdamW
Initial learning rate	10^{-3}
Batch size	6 per GPU, 18 global
Image augmentations	Intensity shift 0.03, intensity scaling 0.05, Gaussian noise $\sigma = 0.01$

Table 1 summarizes the training protocol. Hardware and runtime details for the final training run are reported in Appendix A, Table 7.

The main lesion loss was defined as

$$\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{Dice}} + 2\mathcal{L}_{\text{BCE}}. \tag{2}$$

The total training loss used the auxiliary supervision loss described in Section 2.2:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + 0.25\,\mathcal{L}_{\text{aux}}. \tag{3}$$

4 Results

LETT-NeXt was evaluated in terms of segmentation accuracy, qualitative behavior, and computational efficiency. Quantitative performance was measured on the public validation set and hidden test set using DSC, NSD, and the challenge score. Metrics are reported on a 0–100 scale, and absolute differences are reported in percentage points.

4.1 Quantitative validation results

LETT-NeXt completed all 49 public validation cases without failures or timeouts. Table 2 summarizes public validation and hidden test performance. Compared with the Lite ENSAM competition baseline, LETT-NeXt achieved higher hidden test DSC, NSD, and challenge score. On public validation, LETT-NeXt obtained higher DSC but lower NSD, resulting in a lower overall validation score.

The larger standard deviation for NSD suggests that boundary accuracy varied more than volumetric overlap. This is expected in RECIST-guided segmentation, because the RECIST marker provides lesion position and approximate scale, but not the full 3D lesion boundary.

Table 2: Quantitative evaluation results on the public validation and hidden test sets. The LETT-NeXt validation result uses the inference pipeline with AutoZoom enabled.

Model	Public Validation			Hidden Test		
	DSC(%)	NSD(%)	Score(%)	DSC(%)	NSD(%)	Score(%)
Lite ENSAM	76.10 ± 16.30	78.90 ± 19.10	77.50	60.70	63.60	62.15
LETT-NeXt	79.43 ± 10.09	72.32 ± 16.22	75.87 ± 12.17	73.90	67.30	70.60

4.2 Qualitative results

Figure 3 shows representative validation examples selected by score percentile. Qualitatively, LETT-NeXt usually localized the prompted lesion, but the main visible failure mode was under-segmentation. In difficult cases, the model predicted masks that were too small, especially for small lesions, low-contrast lesions, and lesions with unclear boundaries.

4.3 Runtime and memory

LETT-NeXt satisfied the CPU inference constraints on the 49-case public validation set. Table 3 summarizes validation set runtime and memory use. The hidden test server reported a mean runtime of 20.9 s per case under hidden challenge-server conditions.

Table 3: Runtime and memory use on the 49-case public validation set.

Quantity	**Value**
Mean runtime per case	6.88 ± 3.00 s
Maximum runtime	13.55 s
Mean CPU memory-time	11.01 ± 5.49 GB s
Total CPU memory-time	539.36 GB s
Maximum resident memory	3.57 GB

5 Ablation Studies

Ablation studies evaluated three components of the RECIST-guided lesion segmentation pipeline: backbone architecture, auxiliary anatomy supervision, and adaptive field-of-view expansion with AutoZoom. These controlled ablations used shorter training runs and smaller crops than the final model, so they should be interpreted as component comparisons rather than direct final-model results. Additional setup details are provided in Appendix C.

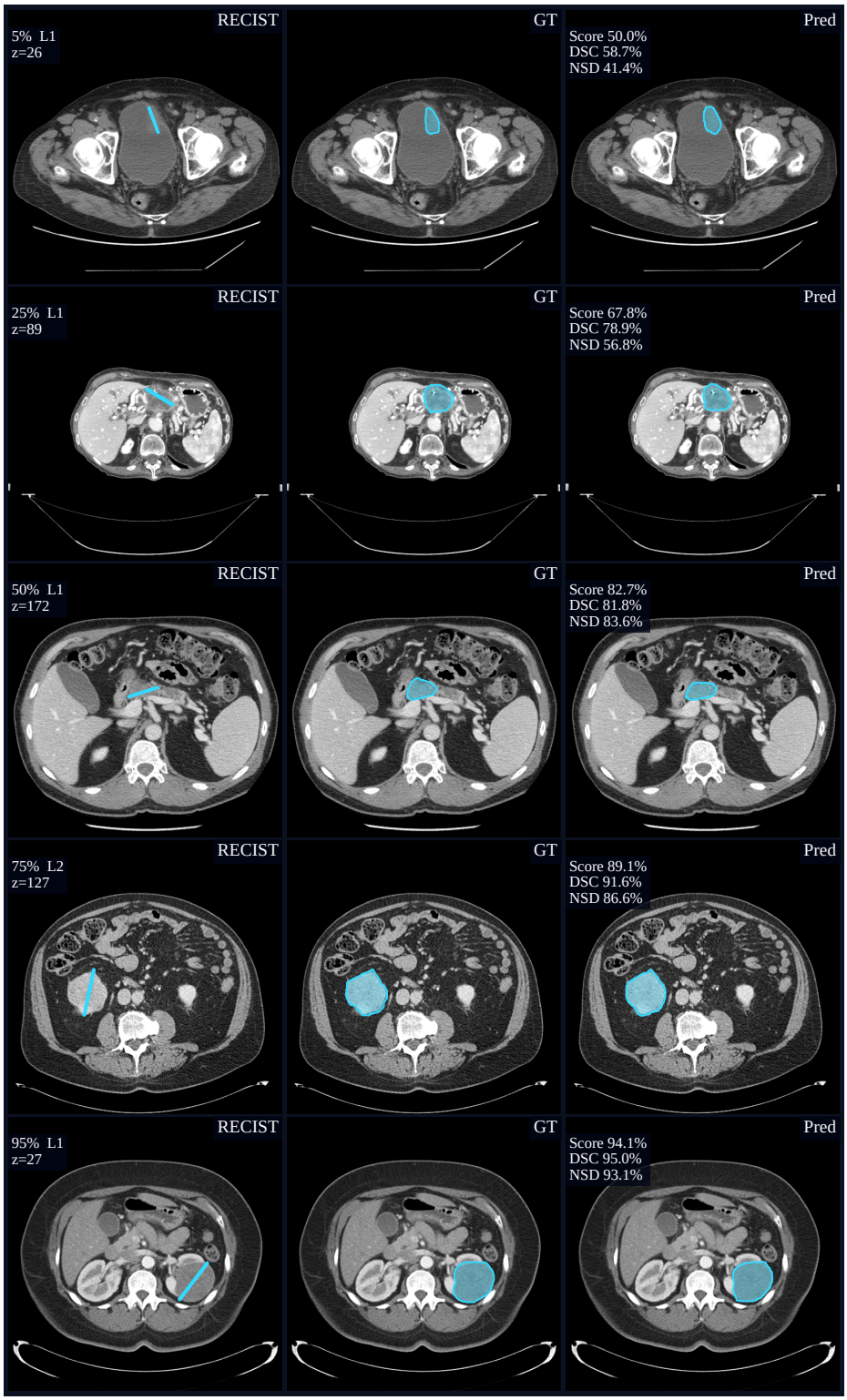

Fig. 3: Qualitative validation examples across score percentiles. Each row shows the RECIST input marker, ground-truth mask, and LETT-NeXt prediction on the prompt slice. Higher-score examples generally show stronger agreement with the ground truth.

5.1 Architecture Ablation

Five segmentation backbones were compared using the controlled protocol described in Appendix C. Auxiliary anatomy supervision was disabled for this comparison.

Table 4 shows the 49-case validation results. MedNeXt-v2 f32 achieved the best validation score, with 71.40, compared with 65.08 for the next-best backbone, SwinUNETR f24. In the paired comparison, this corresponded to an improvement of +6.32 percentage points, with a 95% bootstrap CI of $[+3.90, +8.80]$ percentage points. MedNeXt-v2 f32 was therefore selected as the final segmentation backbone.

Table 4: Matched architecture ablation on the 49-case validation set.

Backbone	**Params.**	**Score (%)**	**DSC (%)**	**NSD (%)**
UNet base	1.19M	63.90	70.87	56.92
SegResNet	4.70M	63.59	72.13	55.06
SwinUNETR f24	15.70M	65.08	72.32	57.84
MedNeXt-v2 small	1.28M	65.04	72.16	57.92
MedNeXt-v2 f32	**6.92M**	**71.40**	**76.06**	**66.74**

5.2 Auxiliary Anatomy Supervision

The auxiliary head was evaluated using a matched 15-epoch control experiment described in Appendix C.3. The comparison tested whether enabling the PanTS auxiliary anatomy–tumor head improved RECIST-guided lesion segmentation.

Table 5 reports the best-threshold validation results. Auxiliary supervision improved the mean validation score by +0.96 percentage points, but the 95% bootstrap confidence interval included zero $[-0.18, +2.05]$. The auxiliary head is therefore interpreted as a weak training-time regularizer rather than a definitive improvement or standalone anatomy segmentation component.

Additional threshold-sweep results are provided in Appendix C.3.

5.3 AutoZoom

Table 6 compares fixed RECIST-centered regional crop inference with adaptive AutoZoom. The table reports the mean challenge score on a 0–100 scale, where Δ denotes AutoZoom minus fixed-crop inference in percentage points. AutoZoom triggered in only a small fraction of cases and had near-neutral aggregate performance. On the validation set, the difference was small, with a 95% bootstrap CI spanning zero $[-0.32, +0.02]$ percentage points. AutoZoom is therefore best

Table 5: Auxiliary-head ablation using a matched MedNeXt-v2 small prompt setup. Metrics are reported as mean ± SD over the 49-case validation set on a 0–100 scale.

Variant	**Threshold**	**Score (%)**	**DSC (%)**	**NSD (%)**
No auxiliary head	0.35	66.13 ± 13.09	72.63 ± 11.18	59.63 ± 18.45
PanTS auxiliary head	0.35	$\mathbf{67.09} \pm \mathbf{13.29}$	$\mathbf{73.39} \pm \mathbf{11.08}$	$\mathbf{60.80} \pm \mathbf{18.68}$
Paired improvement	–	$+0.96 \pm 4.01$	$+0.76 \pm 3.15$	$+1.17 \pm 5.21$

Table 6: AutoZoom ablation. Scores denote the mean challenge score on a 0–100 scale. Δ is reported as AutoZoom minus fixed-crop inference in percentage points. "Largest" refers to the largest PanTS targets by lesion size.

Subset	N	**Triggers**	**Challenge score (%)**		Δ **(pp)**
			Fixed crop	**AutoZoom**	
Largest 10	10	1 (10.0%)	62.46	63.51	+1.05
Largest 40	40	5 (12.5%)	57.03	58.05	+1.02
Largest 120	120	7 (5.8%)	63.91	64.34	+0.43
All PanTS	801	23 (2.9%)	64.10	64.19	+0.09
Validation set	49	3 (6.1%)	75.99	75.87	-0.12

interpreted as a selective fallback for cases where the initial RECIST-centered crop may truncate the target lesion.

Adaptive expansion was retained to improve robustness in context-limited cases rather than to increase average validation performance. The largest-lesion subsets showed small positive changes, while the full validation set showed a slight decrease. These results suggest that AutoZoom can recover missed lesion extent when additional field of view is needed, but may also reduce segmentation quality when the expanded crop lowers the effective spatial resolution.

Figure 4 shows qualitative validation examples. Adaptive expansion recovered missed lesion extent in some cases, but reduced segmentation quality in others.

6 Discussion

LETT-NeXt demonstrates that sparse RECIST markers can support efficient 3D lesion segmentation under CPU inference constraints. The model generally localized the marked lesions, while the remaining errors mainly involved lesion extent and boundary placement. This pattern is consistent with the lower NSD than DSC on the public validation set and with the under-segmentation examples in Figure 3. In practice, this suggests that the RECIST marker provides strong localization information, but does not fully resolve the 3D lesion boundary.

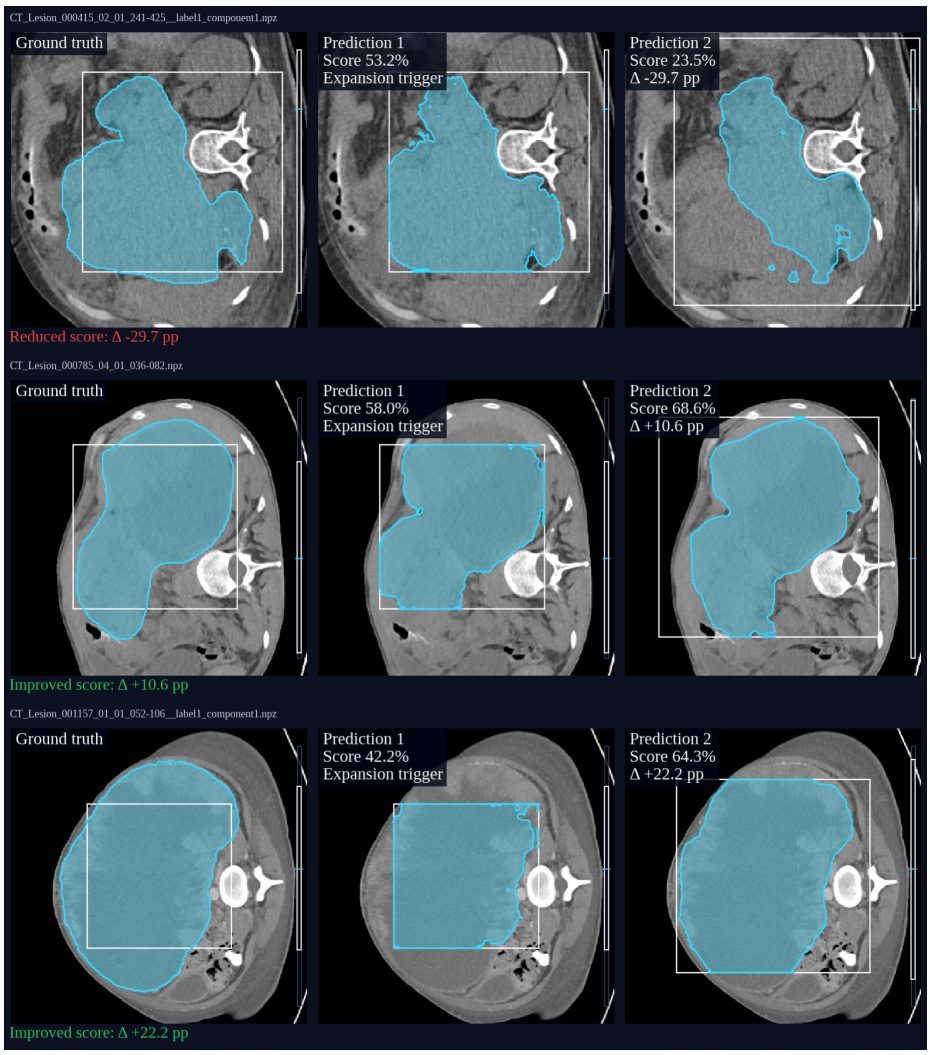

Fig. 4: Qualitative examples of RECIST-guided AutoZoom on the validation set. Each row shows the ground truth, the first prediction using the initial crop, and the second prediction after adaptive crop expansion.

The controlled ablations suggest that backbone choice was the main driver of performance. The MedNeXt-v2 f32 backbone provided the clearest performance gain, whereas auxiliary supervision acted more like weak training-time regularization than a definitive improvement. AutoZoom was neutral on average, but remained useful as a selective fallback for cases where the initial regional crop may not fully contain the lesion.

The main limitations are the small 49-case validation set, threshold selection on the same local validation split used for model development, limited hidden-test detail, and the absence of external validation. The ablations also tested components mostly in isolation, so interactions between backbone choice, auxiliary supervision, crop strategy, and postprocessing remain uncertain. Future work should evaluate LETT-NeXt on external datasets and improve boundary refinement. A promising direction is to pre-train a larger anatomy-aware model on dense multi-class anatomy–tumor labels, for example, from PanTS, before fine-tuning it for RECIST-marker tumor segmentation or distilling it into a compact student model. This strategy could improve anatomical context and boundary modeling while preserving efficient CPU inference.

Overall, LETT-NeXt shows that RECIST markers can support accurate and efficient local 3D lesion segmentation, but further work is needed to improve boundary accuracy and robustness across datasets.

Acknowledgements

The authors express their appreciation to Novartis Norge AS and Akershus University Hospital for funding this work. The submitted method for the *Foundation Models for Pan-cancer Segmentation in CT Images* competition is fully automatic and does not require manual intervention. Only data permitted by the organizers was used. We thank the challenge organizers, data providers, and CodaBench [25] for hosting the competition.

A Implementation Details

Training used distributed data parallelism on three NVIDIA GeForce RTX 5090 GPUs. Table 7 summarizes the hardware and runtime used for the final training run.

B MedNeXt-v2 Building Blocks

Figure 5 summarizes the core MedNeXt-v2 blocks used in LETT-NeXt.

C Controlled Ablation Details

This appendix provides the shared protocol for the controlled ablations reported in Section 5. The architecture ablation and auxiliary-head control used the same

Table 7: Hardware and runtime used for the final training run.

Item	Value
Operating system	Linux 6.1.0-44-amd64
CPU	24 physical cores / 48 threads
System memory	202 GB RAM
GPUs	3 × NVIDIA GeForce RTX 5090
GPU memory	34 GB per GPU
Training parallelism	Distributed data parallel
Data-loader workers	8 per process
Training runtime	25.2 h for the complete training

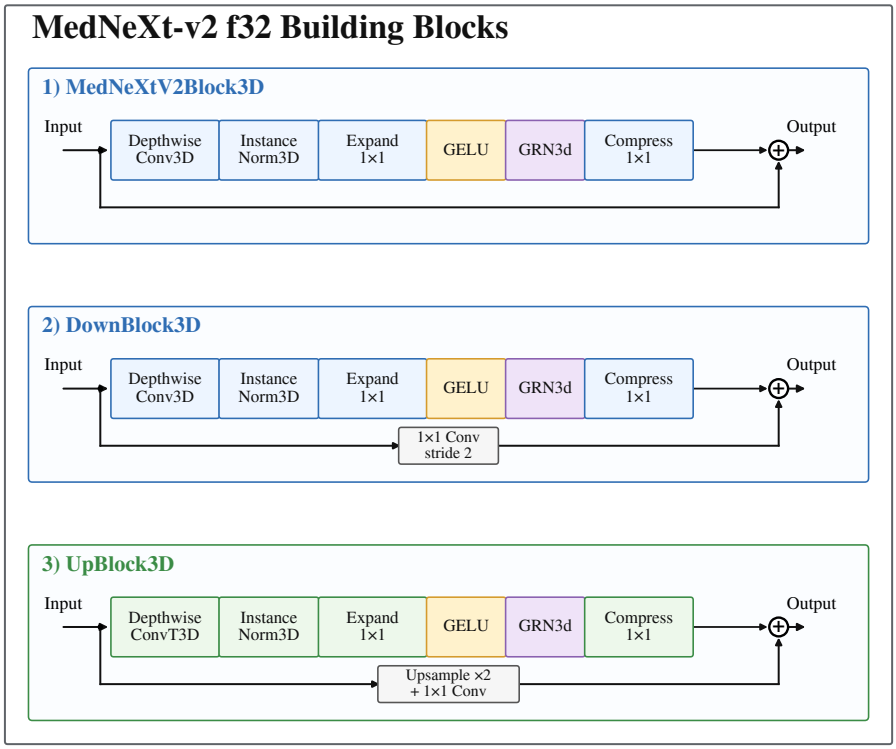

Fig. 5: Core MedNeXt-v2 building blocks used in LETT-NeXt. The main block uses depthwise convolution, normalization, pointwise expansion and compression, GELU activation, global response normalization, and a residual connection. The downsampling and upsampling blocks use residual projections to change spatial resolution.

Table 8: Overview of the backbones included in the architecture ablation.

Model	**Family**	**Params.**
UNet base	3D UNet	1.19 M
SegResNet	Residual CNN	4.70 M
SwinUNETR f24	Transformer–CNN hybrid	15.70 M
MedNeXt-v2 small	ConvNeXt-style 3D CNN	1.28 M
MedNeXt-v2 f32	ConvNeXt-style 3D CNN	6.92 M

training data, crop geometry, prompt encoding, training schedule, validation set, and postprocessing unless otherwise stated. Paired improvements were computed from per-case validation-score differences on the 49-case validation set. Uncertainty was estimated using 95% percentile bootstrap confidence intervals over validation cases.

C.1 Shared Training and Validation Setup

All controlled ablation models were trained for 15 epochs with 625 steps per epoch and a global batch size of 8. This corresponds to 5,000 sampled crop presentations per epoch and 75,000 crop presentations in total. The crop size was $64 \times 128 \times 128$ voxels and the target spacing was $2.4 \times 1.0 \times 1.0$ mm, both given in z, y, x order.

The training data contained 25,112 records: 24,311 primary training records and 801 PanTS records. Validation was performed on the 49-case validation set. All models used the same three-channel input formulation defined in Section 2.1. The main output was a single binary lesion logit. The main lesion loss was Dice loss plus BCE loss with weights 1:2. Validation used prompt-component postprocessing, AutoZoom, threshold 0.5 unless otherwise stated, and cached dataset-statistics normalization.

C.2 Architecture Ablation

The architecture ablation compared five segmentation backbones under the shared setup described in Section C.1. The only intended experimental variable was the backbone architecture. Auxiliary anatomy supervision was disabled for all models in this ablation.

Table 8 summarizes the backbones included in the comparison. All models used the same three-channel input formulation and binary lesion output, keeping the prompt representation fixed across backbones. The RECIST marker and endpoint maps were concatenated with the CT crop, so each backbone received the prompt as voxel-aligned image information.

Table 9: Threshold sweep for the auxiliary-head control experiment. Scores are reported on a 0–100 scale, and differences are reported in percentage points.

Threshold	No aux score (%)	Aux score (%)	Δ (pp)
0.35	66.13	**67.09**	+0.96
0.40	65.65	67.02	+1.37
0.45	65.41	66.78	+1.37
0.50	65.02	66.36	+1.34
0.55	64.84	65.79	+0.95
0.60	64.49	65.45	+0.95
0.65	63.96	64.77	+0.81
0.70	63.34	64.02	+0.69
0.75	62.58	63.32	+0.74

C.3 Auxiliary-Head Control

The auxiliary-head control compared two MedNeXt-v2 small models under the shared setup described in Section C.1. The only intended experimental variable was whether the PanTS auxiliary anatomy–tumor head was enabled during training.

Table 9 reports the threshold sweep for the auxiliary-head control experiment. The auxiliary-head variant achieved a higher validation score at every swept threshold.

The effect of auxiliary supervision was likely limited by its small share of the training data: PanTS contributed 801 of 25,112 training records, corresponding to 3.2%. The auxiliary branch also learned a tumor signal on the 50-case PanTS auxiliary evaluation set, reaching an auxiliary tumor DSC of 62.38%. These results should be interpreted as a matched single-run control rather than a multi-seed estimate.

References

1. FLARE-MedFM/FLARE-Task1-PancancerRECIST-to-3D · Datasets at Hugging Face.
2. Agnar Martin Bjornstad, Elias Stenhede, and Arian Ranjbar. Lite ENSAM: a lightweight cancer segmentation model for 3D Computed Tomography, November 2025. arXiv:2511.01600 [cs] version: 1.
3. Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts, March 2026. arXiv:2511.16719 [cs.CV].

4. E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, January 2009.
5. S.A. Hayes, M.C. Pietanza, D. O'Driscoll, J. Zheng, C.S. Moskowitz, M.G. Kris, and M.S. Ginsberg. Comparison of CT volumetric measurement with RECIST response in patients with lung cancer. *European Journal of Radiology*, 85(3):524–533, March 2016.
6. Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, Daguang Xu, and Wenqi Li. VISTA3D: A Unified Segmentation Foundation Model For 3D Medical Imaging, November 2024. arXiv:2406.05285 [cs].
7. Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, and Yu Qiao. STU-Net: Scalable and Transferable Medical Image Segmentation Models Empowered by Large-Scale Supervised Pre-training, April 2023. arXiv:2304.06716 [cs].
8. Hartmut Häntze, Lina Xu, Christian J. Mertens, Felix J. Dorfner, Leonhard Donle, Felix Busch, Avan Kader, Sebastian Ziegelmayer, Nadine Bayerl, Nassir Navab, Daniel Rueckert, Julia Schnabel, Hugo JWL Aerts, Daniel Truhn, Fabian Bamberg, Jakob Weiß, Christopher L. Schlett, Steffen Ringhof, Thoralf Niendorf, Tobias Pischon, Hans-Ulrich Kauczor, Tobias Nonnenmacher, Thomas Kröncke, Henry Völzke, Jeanette Schulz-Menger, Klaus Maier-Hein, Mathias Prokop, Bram van Ginneken, Alessa Hering, Marcus R. Makowski, Lisa C. Adams, and Keno K. Bressem. MRSegmentator: Multi-Modality Segmentation of 40 Classes in MRI and CT. *Radiology: Artificial Intelligence*, 7(6):e240777, November 2025. arXiv:2405.06463 [eess].
9. Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation, September 2018. arXiv:1809.10486 [cs].
10. Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, Jonathan Deissler, Ralf Floca, and Klaus Maier-Hein. nnInteractive: Redefining 3D Promptable Segmentation, March 2025.
11. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything, April 2023. arXiv:2304.02643 [cs.CV].
12. Wenxuan Li, Xinze Zhou, Qi Chen, Tianyu Lin, Pedro R. A. S. Bassi, Szymon Plotka, Jaroslaw B. Cwikla, Xiaoxi Chen, Chen Ye, Zheren Zhu, Kai Ding, Heng Li, Kang Wang, Yang Yang, Yucheng Tang, Daguang Xu, Alan L. Yuille, and Zongwei Zhou. PanTS: The Pancreatic Tumor Segmentation Dataset, 2025. Version Number: 1.
13. Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s, March 2022. arXiv:2201.03545 [cs].
14. Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment Anything in Medical Images. *Nature Communications*, 15(1):654, January 2024. arXiv:2304.12306 [eess.IV].

15. Jun Ma and Bo Wang, editors. *Fast, Low-Resource, Accurate Robust Organ and Pan-cancer Segmentation: MICCAI Challenge, FLARE 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 6, 2024, Proceedings*, volume 15717 of *Lecture Notes in Computer Science*. Springer Nature Switzerland, Cham, 2026.
16. Léo Machado, Léo Alberge, Hélène Philippe, Elodie Ferreres, Julien Khlaut, Julie Dupuis, Korentin Le Floch, Denis Habip Gatenyo, Pascal Roux, Jules Grégory, Maxime Ronot, Corentin Dancette, Tom Boeken, Daniel Tordjman, Pierre Manceron, and Paul Hérent. A promptable CT foundation model for solid tumor evaluation. *npj Precision Oncology*, 9(1):121, April 2025.
17. Andriy Myronenko. 3D MRI brain tumor segmentation using autoencoder regularization, November 2018. arXiv:1810.11654 [cs].
18. Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos, October 2024. arXiv:2408.00714 [cs.CV].
19. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. arXiv:1505.04597 [cs].
20. Saikat Roy, Yannick Kirchhoff, Constantin Ulrich, Maximillian Rokuss, Tassilo Wald, Fabian Isensee, and Klaus Maier-Hein. MedNeXt-v2: Scaling 3D ConvNeXts for Large-Scale Supervised Representation Learning in Medical Image Segmentation, December 2025. arXiv:2512.17774 [eess].
21. Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jäger, and Klaus H. Maier-Hein. MedNeXt: Transformer-Driven Scaling of ConvNets for Medical Image Segmentation. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 405–415, Cham, 2023. Springer Nature Switzerland.
22. Elias Stenhede, Agnar Martin Bjørnstad, and Arian Ranjbar. ENSAM: an efficient foundation model for interactive segmentation of 3D medical images, September 2025. arXiv:2509.15874 [cs].
23. Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, Bin Fu, Shaoting Zhang, Junjun He, and Yu Qiao. SAM-Med3D: Towards General-purpose Segmentation Models for Volumetric Medical Images, September 2024. arXiv:2310.15161 [cs.CV].
24. Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: robust segmentation of 104 anatomical structures in CT images. *Radiology: Artificial Intelligence*, 5(5):e230024, September 2023. arXiv:2208.05868 [eess].
25. Zhen Xu, Sergio Escalera, Isabelle Guyon, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, and Huan Zhao. Codabench: Flexible, Easy-to-Use and Reproducible Benchmarking Platform, February 2022. arXiv:2110.05802 [cs.LG].