What Matters is the Prompt: Prompt Sensitivity and Prompt Generation in Foundation Models for Lung Nodule Segmentation

Jorge F. Lazo[1], Xixi Liu[1,2], Andreas Hallqvist[3,4], Mikael Johansson[5], Åse Johnsson[6,7], Jonas S. Andersson[5], Jennifer Alvén[1], and Ida Häggström[1,8]

[1] Department of Electrical Engineering, Chalmers University of Technology
[2] Department of Computing, Imperial College London
[3] Department of Oncology, Sahlgrenska University Hospital
[4] Department of Oncology, Institute of Clinical Sciences, University of Gothenburg
[5] Department of Diagnostics and Intervention, Umeå University
[6] Department of Radiology, Sahlgrenska University Hospital
[7] Department of Radiology, Institute of Clinical Sciences, University of Gothenburg
[8] Department of Medical Radiation Sciences, Institute of Clinical Sciences, University of Gothenburg

Abstract. Lung nodule segmentation in computed tomography is essential for extracting clinically relevant information for lung cancer assessment and treatment planning. Foundation models have shown notable segmentation capabilities, but state-of-the-art approaches often depend on input prompts, such as points or boxes, making their performance sensitive to prompt quality and placement. Understanding the limitations and constraints of prompt-based foundation models is therefore essential for designing reliable medical image segmentation solutions. In this work, we investigate how prompt quality affects foundation models performance for lung nodule segmentation. We further propose a synthetic prompt-generation model to test if the dependence on manually provided prompts can be mitigated by generating synthetic prompts that can also improve segmentation performance. Perturbation experiments show that bounding box prompts generally outperform point prompts, while latest specialized medical imaging models achieve better performance than general purpose ones. The proposed approach obtains a Dice coefficient of 0.85, suggesting that synthetic prompt generation as a promising strategy for lung nodule segmentation with foundation models.

Keywords: image segmentation · foundation models · lung cancer · prompt generation

1 Introduction

Lung cancer is one of the deadliest and most common cancers in the world [15]. Computed tomography (CT) is a widely used imaging technique for the preliminary diagnosis of lung cancer and the detection of lung nodules. Early detection is

crucial to improving patient outcomes [4]. Nevertheless, identifying lung nodules is a complex and time-consuming task. Accurate segmentation of lung nodules in CT scans is essential to provide critical information about their size, location, and extent for disease evaluation and treatment planning. However, obtaining precise segmentations of small anatomical structures, such as lung nodules, remains a challenge.

With the rise of AI tools and methods, different models have been proposed to assist clinicians in the task of lung nodule detection and segmentation [13,10]. Recently, foundation models have emerged as a promising paradigm for natural image segmentation, such as the Segment Anything (SAM) [3] model in natural images and MedSAM [7] in the medical imaging domain. Unlike conventional segmentation models trained for a fixed set of object categories, segmentation foundation models are designed to perform segmentation across a wide variety of domains and object types, often through user prompts (e.g., points, bounding boxes, or text), with little or no task-specific fine-tuning. They serve as reusable foundations that can be adapted to numerous downstream segmentation tasks.

In the case of lung nodule segmentation, foundation models have mainly been explored as components within larger classification, detection and segmentation pipelines[2,11,12]. More recent works have attempted to adapt foundation models more specifically to medical CT segmentation: StructSAM[6] introduces structure-aware prompt adaptation to improve robustness for lung cancer lesion segmentation, while Probabilistic SAM extends SAM to produce multiple plausible masks and model annotation ambiguity, which is particularly relevant for lung nodules because boundaries can vary between experts [14]. Other specialized models such as LNTransformer [8]use a fine-tuned variant of SAM to refine proposed segmentation masks given predicted boxes using deformable DETR.

Despite these advantages, current foundation-model-based approaches still present important limitations for lung nodule segmentation. First, most of them remain prompt-dependent: they require points, bounding boxes, text prompts, or previous masks to guide the segmentation. This is a strong assumption in clinical workflows, because such prompts are not always available and may require manual input or a separate detection model. Second, the quality of the final mask is highly sensitive to the quality of the prompt; small changes in the location or size of a bounding box can lead to under-segmentation, over-segmentation, or inclusion of nearby anatomical structures. This problem is especially relevant for pulmonary nodules, which are often small, low-contrast, irregularly shaped, and attached to vessels or pleural surfaces. Third, although generic models such as SAM can segment many natural image objects, their zero-shot performance is usually weaker in medical images, where CT intensity distributions, volumetric context, and subtle lesion boundaries differ substantially from the images used during pretraining.

Therefore, in this work, we address two complementary objectives. First, we investigate the influence of prompt type and prompt quality on segmentation performance of lung nodules. Second, motivated by the observed robustness characteristics, we investigate whether automatically generated coarse prompts

can replace manual interaction, enabling a practical prompt-free segmentation pipeline while retaining the benefits of prompt-based foundation models. Unlike previous approaches, that use an additional network detector to produce external bounding-box prompts for SAM, the proposed model learns coarse prompt proposals in a lightweight network directly from the foundation model encoded features, making the segmentation pipeline compact and tightly coupled to the foundation model.

2 Method

A common characteristic between the SAM and MedSAM models family is that they are composed by an image encoder, a prompt encoder and a mask decoder, and depending on the settings, they can produce either 1 single output mask or propose 3 output masks and their respective confidence score (estimated Intersection Over Union, IoU). In the case of the second versions; SAM2 [9] and MedSAM2 [15] they can perform segmentation on volumetric data, which in the case of CT-scans means the whole 3D CT-volume can be used as input instead of a single 2D image slice, such as in the case of SAM and MedSAM.

We assessed how prompt quality affects mask predictions in SAM, SAM2, MedSAM, and MedSAM2. In the baseline experiments we evaluated the predicted masks generated with each of the models on LUNA16 dataset using 3 different input prompts: point, bounding box and point+bounding box. The center point and the limits of the bounding boxes were extracted directly from the ground truth masks. For each model we tested the different checkpoints available in their official repositories.

For SAM and MedSAM; as well as SAM2 and MedSAM2 in image input modality, only nodule-containing CT slices were used, with each slice repeated three times to match the required 3-channel input. For SAM2 and MedSAM2 in 3D mode, the full CT volume was used. Individual predictions were run per nodule, in the cases where $n > 1$ nodules were present in the image, the final prediction was set to the combination of the n predictions. Single and multi-image outputs were evaluated for each prompt type. Example results are shown in Fig. 1.

2.1 Perturbation studies

To test the robustness of the models towards changes in the input prompt we performed two different perturbation studies. In the first set, we tested the stability of the predictions when the input point (p_{x_0}, p_{y_0}) was shifted within a random radius r such that the shifted point was $(p_{x_n}, p_{y_n}) = (p_{x_0} + r, p_{y_0} + r)$. In the cases of the bounding-box and bounding-box+point as input prompts, the center of the box was also moved to match with the shifted center point.

The second perturbation analysis focused on changing the size of the boxes by l pixels to see if adding more background information, or reduce information about the nodule boundaries would affect the performance of the models. In this

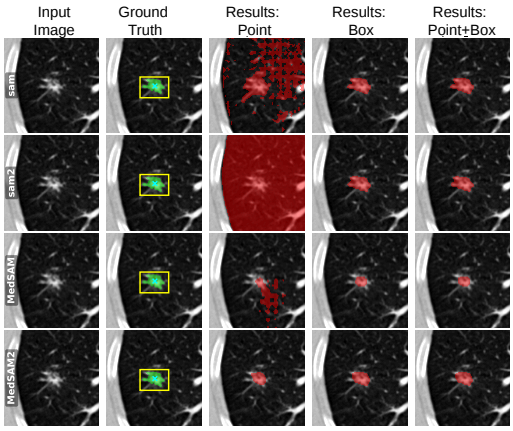

Fig. 1. Samples of the predicted masks by the different foundation models. The input to the models was a 3-channel image as show in the picture, and one of the different combinations of input prompts i.e. point, box or point+box.

case, given an initial bounding box $\{(x_{min}, y_{min}), (x_{max}, y_{max})\}$ that exactly matches the boundaries of the mask, the resized mask was $\{(x_{min} - l, y_{min} - l), (x_{max} + l, y_{max} + l)\}$ with the corresponding version of the points and boxes for the 3D input cases.

2.2 Supervised Decoder Adaptation for Segmentation and Synthetic Prompt Generation

Given that one of the current limitations of image-segmentation foundation models is their reliance on input prompts, we propose adapting these models for segmentation in a fully supervised manner and prompt generation.

Considering $I \in \mathbb{R}^{Z \times H \times W}$ a CT volume, where z indexes the axial slice. For each target slice z, the network input was defined as a three-channel 2.5D image $x_i \in \mathbb{R}^{3 \times H \times W}$ by stacking the neighboring axial slices: $x_i = \{I_{z-1}, I_z, I_{z+1}\}$. In the fully supervised adaptation setup, the pre-trained image encoder E_θ was used to extract dense image embeddings h_i from x_i inputs. The resulting embeddings were passed to an FPN-style decoder [5] where multi-scale features are progressively up-sampled and combined with higher-resolution feature maps before the final mask prediction $\hat{y}_i$. The decoder was trained in a fully supervised manner using annotated nodule masks.

In the synthetic prompt-generation setup, an FPN-based decoder network $D_\varphi(h_i)$ is trained to predict a coarse nodule segmentation mask, and point $\hat{c}_i$ and box $\hat{b}_i$ prompts proposals from the dense feature representations h_i. Point and box predictions are regressed from spatial feature maps formed by combining the refined FPN features and the predicted coarse-mask probability map. The coarse mask is trained with Sørensen-Dice coefficient (DSC) and binary cross-entropy losses against the ground-truth nodule mask. The synthetic point

and box predictions are supervised using normalized L_1 losses for point and box coordinates, together with a generalized IoU loss for the predicted bounding box. In addition, the predicted synthetic prompts are passed through the foundation model prompt encoder $P_\theta(\hat{c}_i, \hat{b}_i)$ and mask decoder to generate prompted-based masks $\{\hat{y}_{Ci}, \hat{y}_{Bi}\}$. The resulting masks are compared with the ground-truth segmentation using DSC losses, encouraging the learned synthetic prompts not only to match geometric targets, but also to maximize the segmentation quality obtained from the prompted foundation model. A schematic of the method is depicted in Fig. 2.

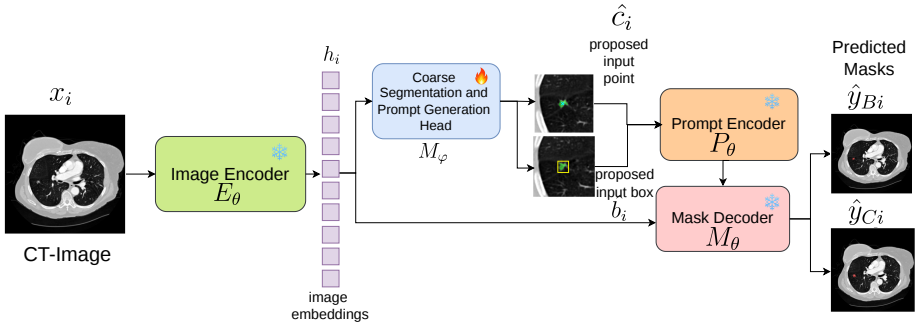

Fig. 2. Prompt generation method. Foundation-model encoder embeddings are used to train the coarse segmentation and synthetic prompt-generation network. The generated points and boxes are passed to the prompt encoder to generate the final masks.

We used the LUNA16 dataset, derived from the LIDC-IDRI dataset [1], which consists of 888 thoracic CT scans containing 1,186 annotated lung nodules, annotated by four radiologists with nodules larger than 3mm considered relevant.

3 Results

The benchmark experiments were performed on the complete LUNA16 dataset. Performance metrics were calculated patient-wise. The mean DSC values were used to report the overall performance of the models in their different input prompts: point, box, point+box; and different output configuration: volume, single image and multi-image. For the multi-output image results, for each prompt type, results are reported for each of the 3 output channels (ch:1, ch:2, ch:3). The results are shown in Table 1.

Overall, box prompts perform better in the single-output image setting. For multi-output images, performance varies by model and channel, but at least one channel consistently performs better. Combining point and box prompts does not generally improve segmentation and can slightly reduce performance in some cases compared with using only a box prompt. This may be because

redundant input information makes the model behave more conservative, causing it to ignore pixels farther from the center point. The performance of point prompts is generally below a DSC of 0.1, except for MedSAM2 which achieves a performance over 0.7. The best results are obtained by SAM2 with a DSC of 0.89 using bounding box prompts with *hierta t* and *hiera s* backbones. Nevertheless, the results for the rest of the 3D models, i.e. SAM2 in its other configurations and MedSAM2 in all of its configurations achieve similar values of 0.88.

Table 1. DSC values of the different foundation models using different input prompts and different output configurations.

model/checkpoint	volume			single image			multi image								
	point	box	point + box	point	box	point + box	point			box			point + box		
							ch:0	ch:1	ch:2	ch:0	ch:1	ch:2	ch:0	ch:1	ch:2
sam/vit_b	–	–	–	0.20	0.86	0.83	0.02	0.03	0.38	0.84	0.86	**0.88**	0.79	0.82	**0.85**
sam/vit_l	–	–	–	0.09	0.87	0.85	0.09	0.28	0.00	0.85	0.87	0.84	0.82	0.86	0.81
sam/vit_h	–	–	–	0.10	0.87	0.85	0.23	0.02	0.00	**0.87**	0.85	0.84	**0.86**	0.83	0.81
medsam/vit_b	–	–	–	0.10	0.79	0.79	0.02	0.04	0.06	0.76	0.75	0.75	0.78	0.77	0.76
medsam/flare22	–	–	–	0.25	0.26	0.25	0.13	0.23	0.28	0.17	0.27	0.30	0.15	0.24	0.27
sam2/hiera_t	0.22	0.33	0.34	0.13	**0.89**	**0.89**	0.01	0.32	0.02	0.46	**0.88**	0.83	0.52	**0.88**	0.83
sam2/hiera_s	0.05	0.08	0.07	0.08	**0.89**	0.88	0.28	0.01	0.00	0.84	0.79	0.46	0.84	0.76	0.34
sam2/hiera_b+	0.19	0.31	0.30	0.10	0.88	0.86	0.02	0.27	0.01	0.50	0.84	0.76	0.45	0.81	0.70
sam2/hiera_l	0.07	0.14	0.15	0.09	0.88	0.87	0.31	0.02	0.01	**0.87**	0.14	0.76	**0.86**	0.09	0.74
medsam2/hiera_t	**0.78**	**0.74**	**0.76**	**0.80**	0.88	0.88	**0.64**	0.73	**0.79**	0.70	0.83	0.84	0.71	0.83	0.84
medsam2/2411	0.77	0.67	0.71	0.74	0.88	0.87	0.33	0.72	0.73	0.76	0.87	0.85	0.76	0.86	0.84
medsam2/CTLesion	0.77	0.66	0.70	0.79	0.88	0.88	0.59	**0.75**	0.77	0.76	0.87	0.85	0.76	0.87	**0.85**

For the perturbation studies, SAM *vit_h*, MedSAM *vit_b*, SAM2 *hierta_t* and MedSAM2 with *hierta_t* backbones were used. In this set of experiments, single-image output and volume configurations were analyzed. In the center point perturbation experiments, all the 3 different input prompts were used, whereas for box-size perturbations, only box and box+point prompts were considered.

Bounding boxes were uniformly perturbed across all coordinates within $-3 \leq l \leq 9$ pixels, with results shown in Fig. 3(a)-(b). In the single image output setting, reducing the box below the nodule size worsens performance, whereas volume output models show the opposite trend. When boxes are expanded beyond the nodule size, performance generally decreases, except for MedSAM, which maintains mean DSC values above 0.7. The results of the center point perturbation are shown in Fig. 3(c)-(e). Center point perturbations were carried out in a range $1 \leq r \leq 9$ pixel-radius away from the original center. MedSAM2 seems to be the more stable model with respect to changes in the center point, and might be related to the way in which it was trained with additional input-prompts from annotators for the same target.

In the Decoder Adaptation (DA) and synthetic prompt generation experiments, the dataset was split patient-wise into training, validation, and test sets using a 0.8/0.1/0.1 ratio. The encoders and decoders of each foundation model were frozen, and only the added segmentation and prompt-proposal heads were trained. For the proposed prompt-generation model, referred to as "*Prompt-*

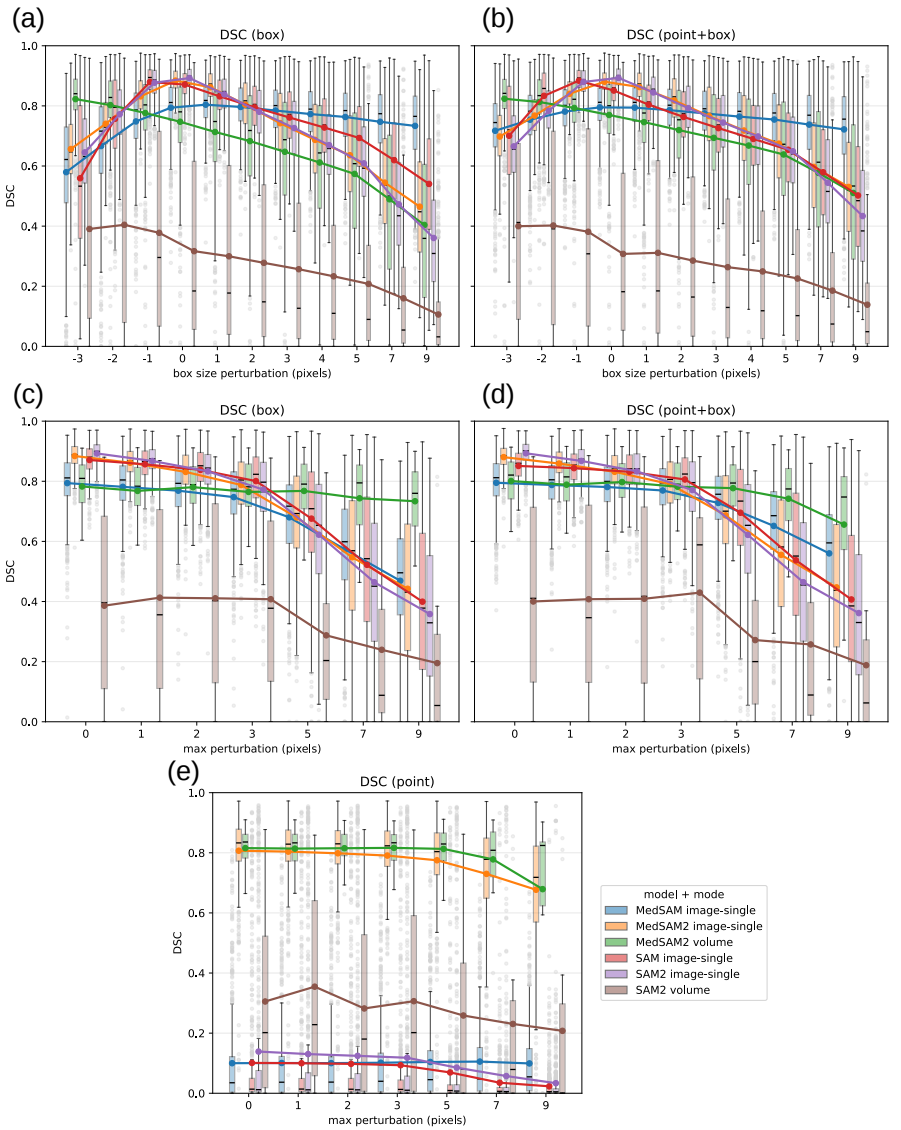

Fig. 3. Perturbation results. (a)-(b) Box plot comparison of the box-size perturbation experiments. (a) using only box as input prompt, (b) combination of point+box. (c)-(e) box plot comparison of the center pixel perturbation experiment according for every input prompt: (c) box, (d) point+box, (e) point.

Gen", the MedSAM2 encoders and decoders were used due to the stability observed in the perturbation studies. The outputs obtained from the generated point and the generated box prompts are reported separately. The proposed method was compared with U-Net and NoduleNet [13], both of which have previously been used for lung nodule segmentation. The results of these experiments are shown in Table 2. All models were trained for 90 epochs using AdamW optimizer with a learning rate of 1×10^{-4}. Performance was evaluated using intersection over union IoU and DSC. The intensities were first windowed by clipping the Hounsfield units to the range [-1000,400] HU. The clipped intensities were then linearly normalized to [0,1]. Intensity augmentation was applied with probability 0.8, including random brightness and contrast changes, additive Gaussian noise, and clipping back to the valid 8-bit range. A Gaussian blur with a 3×3 kernel was additionally applied with probability 0.1. Overall, using SAM2 and MedSAM2 image encoders as backbones led to better results than their SAM and MedSAM counterparts. Prompt generation with the proposed model further improved the results, although the gains were modest.

The results obtained show that foundation models such as SAM2 and MedSAM2 can achieve strong lung nodule segmentation performance when provided with good quality prompts. The proposed Prompt-Gen model represents a step toward reducing the need for manual prompt annotation, but its improvements remain modest and its performance is still constrained by the accuracy of the generated prompts and the limited adaptation of the frozen backbone. These findings suggest that prompt generation is a promising direction when using foundation models in medical image segmentation, but further work is needed to improve robustness, generalization, and end-to-end task adaptation. Future work could incorporate explicit localization losses, uncertainty estimation, or multiple candidate prompts to improve robustness. Moreover, evaluating the method on external datasets would be important to determine whether the observed improvements generalize beyond LUNA16 and across different acquisition protocols, scanners, and annotation styles.

Table 2. Comparison of nodule segmentation models.

Model	DSC	IoU
U-Net	0.781	0.640
NoduleNet	0.826	0.704
DA-SAM	0.692	0.529
DA-MedSAM	0.683	0.518
DA-SAM2	**0.844**	0.720
DA-MedSAM2	0.826	0.703
Prompt-Gen (point)	0.820	0.694
Prompt-Gen (box)	0.843	**0.728**

4 Conclusions

In this work, we analyzed how input prompt quality affects the performance of different foundation models for medical image segmentation. We also proposed two adaptations: using foundation models encoders as frozen backbones, and training a prompt-generation model to automatically produce prompts for mask prediction. The results of the proposed adaptation and prompt generation show that, depending on the backbone, foundation models can improve performance with respect to baseline methods in fully supervised settings. In addition, prompt generation achieved better results than using the models only as frozen backbones, although the improvement was modest. These findings suggest that prompt generation could be a valid strategy for adapting prompt-based foundation models when manual prompts are not available. In general, understanding the limitations of foundation models and the possible strategies for adapting them is essential when designing medical image segmentation solutions. This is particularly relevant in clinical practice, where model sensitivity to prompt quality, prompt placement, and adaptation strategy should be carefully considered before deployment.

Acknowledgments. The study was supported by the Sjöberg Foundation grant 2022-489. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant 2022-06725.

References

1. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)
2. Asha, V., Bhavanishankar, K.: Advanced lung nodule segmentation and classification for early detection of lung cancer using sam and transfer learning. preprint pp. 1–25 (2024)
3. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
4. de Koning, H.J., van Der Aalst, C.M., de Jong, P.A., Scholten, E.T., Nackaerts, K., Heuvelmans, M.A., Lammers, J.W.J., Weenink, C., Yousaf-Khan, U., Horeweg, N., et al.: Reduced lung-cancer mortality with volume ct screening in a randomized trial. New England journal of medicine **382**(6), 503–513 (2020)
5. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
6. Liu, M., Yao, Y., Jia, J., Yao, J., Huang, Z., Zeng, Z., Pu, G., Wu, Y., Bai, Y., Wang, B., et al.: Structsam: structure-aware prompt adaptation for robust lung cancer lesion segmentation in ct. npj Digital Medicine (2026)

7. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
8. Ramezani, H., Vedrines, C., Aleman, D., Létourneau, D.: Lntransformer: Lung nodule transformer for sparse ct segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4238–4245 (2025)
9. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations. vol. 2025, pp. 28085–28128 (2025)
10. Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis **42**, 1–13 (2017)
11. Shaukat, F., Anwar, S.M., Parida, A., Lam, V.K., Linguraru, M.G., Shah, M.: Lung-cadex: Fully automatic zero-shot detection and classification of lung nodules in thoracic ct images. In: International Workshop on Machine Learning in Medical Imaging. pp. 73–82. Springer (2024)
12. Swinburne, N.C., Jackson, C.B., Pagano, A.M., Stember, J.N., Schefflein, J., Marinelli, B., Panyam, P.K., Autz, A., Chopra, M.S., Holodny, A.I., et al.: Foundational segmentation models and clinical data mining enable accurate computer vision for lung cancer. Journal of Imaging Informatics in Medicine **38**(3), 1552–1562 (2025)
13. Tang, H., Zhang, C., Xie, X.: Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 266–274. Springer (2019)
14. Ward, T., Imran, A.: A probabilistic segment anything model for ambiguity-aware medical image segmentation. In: Medical Imaging 2026: Imaging Informatics. vol. 13930, pp. 7–12. SPIE (2026)
15. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)