Interactive Segmentation and Report Generation for CT Images

Yannian Gu[1,*], Wenhui Lei[1,*], Hanyu Chen[2], Shaoting Zhang[1], Xiaofan Zhang[1,†]

[1]Shanghai Jiao Tong University, Shanghai, China
[2]China Medical University, Shenyang, China

[*]Contribution Equally. [†]Corresponding Author.

Abstract. Automated CT report generation plays a crucial role in improving diagnostic accuracy and clinical workflow efficiency. However, existing methods lack interpretability and impede patient-clinician understanding, while their static nature restricts radiologists from dynamically adjusting assessments during image review. Inspired by interactive segmentation techniques, we propose a novel interactive framework for 3D lesion morphology reporting that seamlessly generates segmentation masks with comprehensive attribute descriptions, enabling clinicians to generate detailed lesion profiles for enhanced diagnostic assessment. To our best knowledge, we are the first to integrate the interactive segmentation and structured reports in 3D CT medical images. Experimental results across 15 lesion types demonstrate the effectiveness of our approach in providing a more comprehensive and reliable reporting system for lesion segmentation and capturing. The source code is publicly available at https://github.com/yanniangu/ISRG-CT-MICCAI2025.

Keywords: Interactive Framework · Segmentation and Report Generation.

1 Introduction

CT (Computed Tomography) serves as an essential diagnostic tool, with radiological reports serving as the primary medium for communicating diagnostic findings to clinicians [13,22]. In recent years, CT report generation has gained significant attention, with advancements focusing on automating the clinical text generation process based on image analysis results [4,7,8,19]. Researchers have developed various approaches to generate reports from CT images, include traditional vision-based methods [10], vision-language models [3,26], and other methods [15,24]. However, current CT report generation methods **lack interpretability** and impede patient-clinician understanding, while offering no interactivity [16]. This produces reports that fail to explain findings clearly or adapt to new clinical cases, resulting in communication gaps and incomplete assessments.

To address these limitations, interactive segmentation techniques have shown promising capabilities in related medical imaging analysis. Models like the Segment Anything Model (SAM) [12,20] enable clinicians to dynamically guide the segmentation process through intuitive interactions, providing immediate visual feedback and greater control over results. These techniques [11,17,18] successfully balance automation with expert input, allowing for precise refinements while maintaining workflow efficiency. However, current interactive segmentation methods lack integrated report generation capabilities, forcing radiologists to manually convert visual findings into structured clinical reports.

In this work, we introduce the first interactive framework for lesion morphology reporting, integrating both intuitive visual segmentation and quantitative textual attribute description for a comprehensive lesion characterization. Our system enables radiologists to generate detailed clinical reports proposed by Lei et al. [14], through minimal point-based interactions during image review, significantly reducing interpretation time while preserving diagnostic precision. By simultaneously generating accurate segmentation results with structured reports, our approach provides improved explanatory capabilities, creating direct visual-textual correspondence between report descriptions and anatomical features in CT images. Additionally, this integrated framework demonstrates considerable potential for facilitating multi-modal medical dataset annotation.

Technically, our approach delivers two key innovations: (1) feature-space clustering-based point refinement that amplifies the impact of expert interactions, and (2) inter-task feature synergy that enables bidirectional information flow between segmentation and attribute description generation. The interactive nature of our system significantly enhances zero-shot performance, allowing radiologists to provide real-time feedback that helps the model adapt to novel lesion types not encountered during training. Our contributions are as follows:

- We propose the interactive framework for joint lesion segmentation and attribute description that integrates point refinement and feature synergy, enabling comprehensive lesion profiling with minimal radiologist interaction;
- Our approach has the potential to enhance clinical workflow efficiency by reducing interpretation time while enhancing diagnostic communication through direct visual-textual correspondence between reports and segmentations;
- Extensive experiments demonstrate superior performance in both segmentation accuracy and attribute description quality, including robust zero-shot capabilities for novel lesion types.

2 Methodology

Our proposed interactive framework enables radiologists to generate detailed lesion profiles through minimal point interactions. As shown in Fig. 1, the model consists of two main components: the *Vision-Prompt Hybrid Token Encode* module and the *Synergistic Task Integration Decode* module. Below, we describe each module in detail.

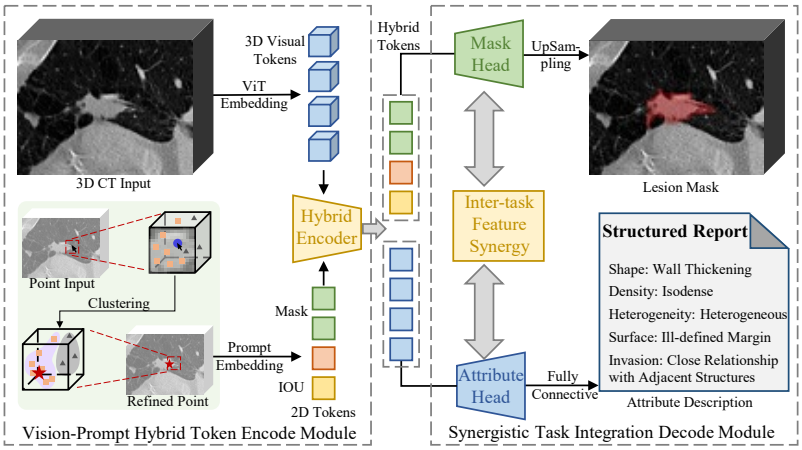

Fig. 1. The model architecture. Our framework processes 3D CT images and user-provided points to simultaneously generate lesion segmentation masks and structured attribute descriptions. The model integrates visual tokens, point tokens, initial mask tokens, and initial IOU tokens into a hybrid encoding module, which forms the foundation for both visual and textual outputs. Key innovations include a clustering-based point refinement technique that optimizes the input points through clustering centers, and an inter-task feature synergy mechanism that enhances the performance of both segmentation and attribute description tasks concurrently.

2.1 Vision-Prompt Hybrid Token Encode Module

This module follows a dual-branch architecture, with image tokens extracted by the *vision encoder* and prompt tokens derived from the *prompt encoder*. Notably, to enhance the accuracy and reliability of user clicks, the prompt encoder integrates a feature-space clustering-based point refinement method. The refined prompt tokens are then combined with the image tokens in the *hybrid token encoder*, producing hybrid tokens that drive subsequent processing.

Vision encoder E_I adopts a ViT-style [5] architecture, consisting of a standard ViT with primary local window attention and several interleaved global attention layers. This design produces isotropic feature maps with consistent feature dimensions. The extracted image features $\mathbf{I} = \mathrm{E}_I(\mathcal{I})$, where $\mathcal{I}$ denotes the input image, are down-sampled for computational efficiency.

Prompt encoder E_P transforms point coordinates using positional embeddings, generating distinct representations based on point labels. This process converts sparse point prompts into embeddings that guide the model's attention toward specific regions of interest. To compensate for the randomness and potential errors in individual point placements and improve interaction efficiency, we incorporate feature-space clustering that extends the influence of each point beyond its immediate location. For each user click p, we define a square local feature window $\mathbf{N}(p)$ of size $p \times p$ centered at the click, and obtain features from

the image features $\mathbf{I}$:

$$\mathbf{I}_p = \{\mathbf{I}(q) | q \in \mathbf{N}(p)\}. \tag{1}$$

Applying K-means clustering, we partition $\mathbf{I}_p$ into k clusters and select the point closest to each centroid μ_i as guidance:

$$q_i^* = \arg\min_{q \in C_i} \|\mathbf{I}(q) - \mu_i\|, i = 1, \cdots, k. \tag{2}$$

These selected points provide structured feedback, enabling the model to refine predictions across semantically related regions rather than isolated pixels. The final point prompt tokens are obtained by $\mathbf{P} = \mathrm{E}_P(q^*)$.

Hybrid token encoder E_H combines image tokens $\mathbf{I}$, prompt tokens $\mathbf{P}$, mask tokens $\mathbf{M}$, and $\mathbf{IOU}$ tokens via self-attention and cross-attention mechanisms, producing the hybrid token $\mathbf{Z} = \mathrm{E}_H(\mathbf{I}, \mathbf{P}, \mathbf{M}, \mathbf{IOU})$. Self-attention aggregates prompt context while cross-attention enables interaction with image information. The mask tokens $\mathbf{M}$ help initialize and guide the lesion segmentation process, while the $\mathbf{IOU}$ tokens provide an initial metric for optimizing the segmentation through the intersection-over-union score.

2.2 Synergistic Task Integration Decode Module

After obtaining the hybrid tokens, we use the *mask head* and the *attribute head* to generate the segmentation masks and attribute descriptions, respectively. Recognizing the intrinsic relationship between these tasks, we introduce the *inter-task feature synergy* mechanism that enables bidirectional information flow, allowing both tasks to mutually enhance each other's performance.

Mask head H_{seg}: Initial mask tokens $\mathbf{M}_0$ interact with the hybrid token $\mathbf{Z}$ through multiple Transformer layers to capture task-specific spatial and semantic features, formulated as $\mathbf{M} = \mathrm{H}_{seg}(\mathbf{M}_0, \mathbf{Z})$. This contextual refinement process allows the mask tokens to encode detailed structural information about the target regions. The extracted mask tokens $\mathbf{M}$ are subsequently upsampled to enhance spatial resolution: $\hat{\mathbf{M}} = \mathrm{Upsample}(\mathbf{M})$, producing the final segmentation mask with precise boundary delineation.

Attribute head H_{attr}: This component predicts the attributes of the segmented regions by processing the shared hybrid token $\mathbf{Z}$. First, $\mathbf{Z}$ is passed through residual block to capture essential features. Next, a self-attention mechanism refines the attribute representations, yielding $\mathbf{A} = \mathrm{H}_{attr}(\mathbf{Z})$. By modeling both local and global dependencies, this architecture enhances the accuracy of attribute predictions. Finally, a classification layer with sigmoid activation generates the multi-label outputs, $\hat{\mathbf{Y}} = \mathrm{FCN}(\mathbf{A})$.

Inter-task feature synergy: Medical image segmentation and attribute prediction are inherently complementary: morphological features provide strong indicators for pathological attributes, while attribute knowledge aids in precise boundary delineation. To leverage this natural synergy, we enable bidirectional feature exchange between the segmentation features $\mathbf{M}$ and attribute features $\mathbf{A}$ by projecting them into a common latent space:

$$\mathbf{Z}_{seg} = \mathrm{Attn}(\mathbf{M}, \mathrm{Proj}(\mathbf{A})), \quad \mathbf{Z}_{attr} = \mathrm{Attn}(\mathbf{A}, \mathrm{Proj}(\mathbf{M})), \tag{3}$$

where $\text{Attn}(\cdot,\cdot)$ represents the cross-attention mechanism that facilitates task-specific features by leveraging complementary information between tasks. The projection operation $\text{Proj}(\cdot,\cdot)$ ensures that the features from both tasks are aligned in the same feature space. This bidirectional interaction improves segmentation accuracy and enhances attribute prediction through mutual guidance.

2.3 Training Strategy

Our training strategy addresses two key challenges: (1) generating accurate segmentation masks $\hat{\mathbf{M}}$ and (2) predicting correct multi-label attributes $\hat{\mathbf{Y}}$. We employ a simulation-based approach to mimic expert guidance and a composite loss function to optimize both objectives simultaneously.

Expert-provided Click Simulation: Expert annotations provide crucial feedback by highlighting misclassified regions and guiding the model to improve boundary delineation and feature representation, but manual corrections are costly. To emulate expert feedback during training, we identify misclassified regions by computing error maps between the predicted mask $\mathbf{M}_{pred}$ and the ground truth $\mathbf{M}_{gt}$:

$$\mathbf{M}_{fn} = \mathbf{M}_{gt} \wedge \neg\mathbf{M}_{pred}, \quad \mathbf{M}_{fp} = \neg\mathbf{M}_{gt} \wedge \mathbf{M}_{pred}. \tag{4}$$

False negatives $\mathbf{M}_{fn}$ represent missed features, while false positives $\mathbf{M}_{fp}$ show incorrect predictions. We randomly select a point from these regions as a simulated expert click, guiding the model's attention to improve performance in subsequent iterations.

Dual-objective Loss Function: We optimize our model using a composite loss function that balances segmentation accuracy and attribute classification: $\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda\mathcal{L}_{attr}$, where λ is an adaptive weighting parameter. For the segmentation component, we employ Dice loss to measure overlap between predicted $\hat{\mathbf{M}}$ and ground truth masks $\mathbf{M}_{gt}$:

$$\mathcal{L}_{seg} = 1 - \frac{2\sum_i(\hat{m}(i) \cdot m_{gt}(i))}{\sum_i \hat{m}(i) + \sum_i m_{gt}(i) + \epsilon}, \tag{5}$$

where $\hat{m}(i)$ and $m_{gt}(i)$ represent the predicted and ground-truth values at voxel i, and ϵ is a small constant to avoid division by zero. For attribute prediction, we use categorical cross-entropy loss:

$$\mathcal{L}_{attr} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_c(n)\log\hat{y}_c(n), \tag{6}$$

where $\hat{\mathbf{Y}} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_C\}$ represents predicted probabilities across C classes, and $\mathbf{Y}gt$ contains one-hot encoded ground-truth labels.

3 Experiments

In this section, we present a comprehensive evaluation of our approach through a series of experiments designed to assess both segmentation performance and structured report generation capabilities.

Table 1. Templates of Structured Lesion Reports. This table summarizes the five key radiological attributes used to characterize lesions, which represents a distinct aspect of lesion morphology and appearance.

Attribute	Description
Shape	"Round-like", "Irregular", "Wall thickening", "Punctate, nodular"
Invasion	"No close relationship with surrounding structures", "Close relationship with adjacent structures"
Density	"Hypodense", "Isodense", "Hyperdence"
Heterogeneity	"Homogeneous", "Heterogeneous"
Surface	"Well-defined margin", "Ill-defined margin"

Table 2. Segmentation Performance. Our proposed method is compared against state-of-the-art segmentation approaches and ablation variants of our model.

Setting	Method	Metrics	
		DSC ↑	**HD95 ↓**
Test	Baseline (UNet) [21]	0.733	5.586
	SegMamba [25]	0.747	5.601
	SAM-Med3D [23]	0.752	5.398
	Ours (Vanilla)	0.758	5.020
	+ Point Refinement	0.764	5.122
	+ Feature Synergy	0.766	4.928
	Ours (Full)	**0.794**	**4.303**
Zero-Shot	Baseline (UNet)	0.712	2.758
	SegMamba	0.712	2.695
	SAM-Med3D	0.718	2.169
	Ours (Vanilla)	0.726	2.079
	+ Point Refinement	0.731	2.018
	+ Feature Synergy	0.727	2.113
	Ours (Full)	**0.735**	**2.001**

3.1 Dataset Description

We assemble 1535 CT scans and masks collected from public and private sources: KiTS23 [9]: kidney tumor and cyst (489 scans); MSD [1]: colon tumor (126 scans), liver tumor (303 scans), lung tumor (96 scans), pancreas tumor (216 scans), pancreas cyst (65 scans); private data collected at China Medical University: liver cyst (30 scans), gallbladder cancer (30 scans), gallstones (30 scans), esophageal cancer (30 scans), gastric cancer (30 scans), kidney stone (30 scans), bladder cancer (30 scans), and bone metastasis (30 scans). All these scans are annotated with structured lesion reports by four radiologists. The gallstone and liver cyst are treated as zero-shot test cases, meaning they are excluded from the training phase. The remaining data is divided into training (60%), validation (20%), and test (20%) sets.

Structured Report: We adopt a structured lesion report template, as proposed in [14]. Specifically, each lesion comes with a corresponding structured textual report, including *shape*, *invasion*, *density*, *heterogeneity*, and *surface*, with multiple possible categories for each. Further details can be found in Table 1.

Data Preprocessing: In the pre-processingg phase, we first locate the largest lesion in each label file to determine its center. This center is then used to crop the CT volume, ensuring that the region of interest (ROI) is tightly focused on the lesion. By isolating the lesion and eliminating irrelevant back-

Table 3. Performance comparison of different models on lesion structured report generation across multiple features, with accuracy as our evaluation metric.

		Shape	Density	Invasion	Surface Invasion	Average	
Test	Baseline(CNN)	0.582	0.562	0.626	0.721	0.716	0.642
	CT-CLIP [6]	0.681	0.525	0.769	0.759	0.746	0.696
	M3D [2]	0.445	0.141	0.488	0.760	0.265	0.420
	Ours (Vanilla)	0.680	0.605	0.769	0.796	0.752	0.720
	+ Point Refinement	0.680	**0.656**	0.782	0.807	**0.762**	0.738
	+ Feature Synergy	0.712	0.623	0.784	**0.850**	0.756	0.745
	Ours (Full)	**0.745**	0.626	**0.833**	**0.850**	0.741	**0.759**
Zero-Shot	Baseline(CNN)	0.283	0.367	0.400	0.167	0.800	0.403
	CT-CLIP	0.283	0.500	0.317	0.000	**1.000**	0.420
	M3D	0.267	0.500	**0.700**	0.000	0.000	0.293
	Ours (Vanilla)	0.283	0.500	0.317	0.317	**1.000**	0.483
	+ Point Refinement	0.283	0.483	0.317	**0.600**	**1.000**	0.537
	+ Feature Synergy	0.283	0.500	0.417	0.500	**1.000**	0.540
	Ours (Full)	**0.300**	**0.600**	0.483	0.533	**1.000**	**0.583**

ground, the model can focus on the lesion's key features, ultimately improving both the efficiency and accuracy of the learning process.

3.2 Comparision with the SOTA Methods

Our evaluation is divided into two key components—segmentation performance and structured attribute description. All compared methods are either trained or fine-tuned on our datasets to ensure a fair comparison. **Ablation Setup**: We evaluate four model variants to isolate component contributions: (1) Vanilla (backbone only), (2) Point Refinement (backbone + feature-space clustering point refinement), (3) Feature Synergy (backbone + Inter-task feature synergy), and (4) our full model combining both enhancements with the backbone.

Segmentation Performance: We compare our model against three leading segmentation methods: *UNet* [21] (the well-established baseline in medical image segmentation), *SegMamba* [25] (which leverages state space models for sequence modeling in volumetric data), and *SAM-Med3D* [23] (a medical adaptation of the Segment Anything Model with 3D capabilities). Performance is evaluated using two metrics: Dice Similarity Coefficient (*DSC*), which measures volumetric overlap, and 95% Hausdorff Distance (*HD95*), which captures boundary precision.

As demonstrated in Table 2, our approach outperforms the compared methods in both standard test cases and more challenging zero-shot scenarios. These results underscore our method's robustness in handling anatomical variability while maintaining precise boundary delineation. Specifically, the Vanilla model's

strong performance confirms the synergy between segmentation and attribute prediction tasks. The Point Refinement enhancement demonstrates how focusing on diagnostically relevant regions improves results, while Feature Synergy validates our approach of connecting spatial and semantic features.

Structured Attribute Description: Structured attribute description can be framed both as a multi-label classification task, where each lesion has multiple associated attributes, and as a visual-language task, where the model answers questions about the lesion's characteristics or aligns the image with corresponding text reports. To evaluate our approach, we compare it against methods from both perspectives with methods *CNN model*, *M3D* [2], and *CT-CLIP* [6]. This allows us to benchmark our method against both specialized VLMs and traditional visual models in the medical vision domain.

As seen in Table 3, our method outperforms others in attribute prediction. The M3D VQA-based approach struggles with complex features (0.000 accuracy for zero-shot surface characteristics), while the CNN baseline falters with lesion boundaries (accuracy drops from 0.642 to 0.403 in zero-shot). CT-CLIP performs well in standard tests but struggles with unseen lesions and morphological features in zero-shot scenarios. Our combined segmentation-attribute approach ensures precise spatial-semantic mapping, yielding superior performance.

3.3 Interactive Framework Performance

To evaluate our interactive framework, we compare against SAM-Med3D, focusing on zero-shot cases. Fig. 2 illustrates qualitative results across click iterations. Our method achieves better boundary delineation and lesion characterization than SAM-Med3D, with progressive improvements as click iterations increase, confirming its ability to refine reports interactively based on radiologist input.

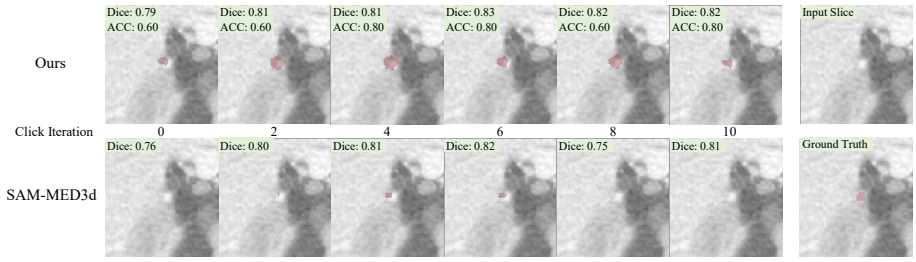

Fig. 2. Qualitative comparison between our method and SAM-Med3D, showing segmentation progression across click iterations. Red overlays indicate masks.

4 Conclusion

We propose an interactive framework for lesion morphology reporting that integrates visual segmentation with structured attribute description. Our approach

enhances segmentation accuracy, improves attribute description precision, strengthens zero-shot performance, and allows radiologists to refine reports interactively. Experiments demonstrate that our method outperforms state-of-the-art approaches, offering a more flexible and accurate solution.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62301311.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
2. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models (2024)
3. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., et al.: Merlin: A vision language foundation model for 3d computed tomography. Research Square pp. rs–3 (2024)
4. Dayarathna, S., Islam, K.T., Uribe, S., Yang, G., Hayat, M., Chen, Z.: Deep learning based synthesis of mri, ct and pet: Review and analysis. Medical image analysis **92**, 103046 (2024)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
6. Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasdelen, M.F., Durugol, O.F., Wittmann, B., Amiranashvili, T., et al.: Developing generalist foundation models from a multimodal dataset for 3d computed tomography. arXiv preprint arXiv:2403.17834 (2024)
7. Hamamci, I.E., Er, S., Menze, B.: Ct2rep: Automated radiology report generation for 3d medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 476–486. Springer (2024)
8. Hamamci, I.E., Er, S., Sekuboyina, A., Simsar, E., Tezcan, A., Simsek, A.G., Esirgun, S.N., Almas, F., Doğan, I., Dasdelen, M.F., et al.: Generatect: Text-conditional generation of 3d chest ct volumes. In: European Conference on Computer Vision. pp. 126–143. Springer (2024)
9. Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao, Z., Chen, H., Wang, L., Golts, A., Khapun, D., Shats, D., Shoshan, Y., Gilboa-Solomon, F., George, Y., Yang, X., Zhang, J., Zhang, J., Xia, Y., Wu, M., Liu, Z., Walczak, E., McSweeney, S., Vasdev, R., Hornung, C., Solaiman, R., Schoephoerster, J., Abernathy, B., Wu, D., Abdulkadir, S., Byun, B., Spriggs, J., Struyk, G., Austin, A., Simpson, B., Hagstrom, M., Virnig, S., French, J., Venkatesh, N., Chan, S., Moore, K., Jacobsen, A., Austin, S., Austin, M., Regmi, S., Papanikolopoulos, N., Weight, C.: The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct (2023)

10. Hou, Z., Yan, R., Yan, Z., Lang, N., Zhou, X.: Energy-based controllable radiology report generation with medical knowledge. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 240–250. Springer (2024)
11. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
13. Lai, Y., Chen, X., Wang, A., Yuille, A., Zhou, Z.: From pixel to cancer: Cellular automata in computed tomography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 36–46. Springer (2024)
14. Lei, W., Chen, H., Zhang, Z., Luo, L., Xiao, Q., Gu, Y., Gao, P., Jiang, Y., Wang, C., Wu, G., et al.: A data-efficient pan-tumor foundation model for oncology ct interpretation. arXiv preprint arXiv:2502.06171 (2025)
15. Li, Y., Wang, Z., Liu, Y., Wang, L., Liu, L., Zhou, L.: Kargen: Knowledge-enhanced automated radiology report generation using large language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 382–392. Springer (2024)
16. Luo, L., Vairavamurthy, J., Zhang, X., Kumar, A., Ter-Oganesyan, R.R., Schroff, S.T., Shilo, D., Hossain, R., Moritz, M., Rajpurkar, P.: Rexplain: Translating radiology into patient-friendly video reports. arXiv preprint arXiv:2410.00441 (2024)
17. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
18. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Medical Image Analysis **89**, 102918 (2023)
19. Nakaura, T., Yoshida, N., Kobayashi, N., Shiraishi, K., Nagayama, Y., Uetani, H., Kidoh, M., Hokamura, M., Funama, Y., Hirai, T.: Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. Japanese Journal of Radiology **42**(2), 190–200 (2024)
20. 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. arXiv preprint arXiv:2408.00714 (2024)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
22. Wang, C., Shao, J., He, Y., Wu, J., Liu, X., Yang, L., Wei, Y., Zhou, X.S., Zhan, Y., Shi, F., et al.: Data-driven risk stratification and precision management of pulmonary nodules detected on chest computed tomography. Nature Medicine **30**(11), 3184–3195 (2024)
23. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d: towards general-purpose segmentation models for volumetric medical images. arXiv preprint arXiv:2310.15161 (2023)
24. Xiang, Z., Cui, S., Shang, C., Jiang, J., Zhang, L.: Gmod: Graph-driven momentum distillation framework with active perception of disease severity for radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 295–305. Springer (2024)

25. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 578–588. Springer (2024)
26. Yin, H., Zhou, S., Wang, P., Wu, Z., Hao, Y.: KIA: Knowledge-guided implicit vision-language alignment for chest X-ray report generation. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics. pp. 4096–4108. Association for Computational Linguistics, Abu Dhabi, UAE (Jan 2025), https://aclanthology.org/2025.coling-main.276/