nature communications

Article

https://doi.org/10.1038/s41467-024-53387-y

LLM-driven multimodal target volume contouring in radiation oncology

Received: 29 October 2023

Accepted: 10 October 2024

Published online: 24 October 2024

Check for updates

Yujin Oh[1,8], Sangjoon Park[2,3,8], Hwa Kyung Byun[4], Yeona Cho[5], Ik Jae Lee[2], Jin Sung Kim[2,6] ✉ & Jong Chul Ye[7] ✉

Target volume contouring for radiation therapy is considered significantly more challenging than the normal organ segmentation tasks as it necessitates the utilization of both image and text-based clinical information. Inspired by the recent advancement of large language models (LLMs) that can facilitate the integration of the textural information and images, here we present an LLM-driven multimodal artificial intelligence (AI), namely LLMSeg, that utilizes the clinical information and is applicable to the challenging task of 3-dimensional context-aware target volume delineation for radiation oncology. We validate our proposed LLMSeg within the context of breast cancer radiotherapy using external validation and data-insufficient environments, which attributes highly conducive to real-world applications. We demonstrate that the proposed multimodal LLMSeg exhibits markedly improved performance compared to conventional unimodal AI models, particularly exhibiting robust generalization performance and data-efficiency.

Despite the rapid development of Artificial Intelligence (AI) models, there is yet a discernible gap in the realm of medical data processing. Historically, AI models have predominantly focused on individual data modalities—either visual or linguistic. This approach starkly contrasts with the intrinsic multimodal practices of physicians, who inherently rely on a confluence of imaging studies and textual electronic medical data for informed decision-making. By understanding diverse data types and their interrelationships, multimodal AIs would facilitate more accurate diagnoses, personalized treatment development, and a reduction in medical errors by providing a comprehensive view of patient data. For example, in the field of radiation oncology, which is one of the clinical fields to evaluate the potential of multimodal AI applications and the main focus of this article, the integration of multiple modalities holds great importance[1].

For modern intensity-modulated radiation therapy and its inverse planning, two critical components are needed: organs-at-risk (OARs) and the target volume where the dose is prescribed. OARs are defined as the radiosensitive organs susceptible to damage by ionizing radiation during radiation therapy. Traditionally, they were either manually delineated by human experts or automatically contoured using atlas-based autocontouring algorithms. However, with the advent of deep learning-based AI models, such tasks have been efficiently accomplished[2,3]. Therefore, these OARs can be contoured "as they appear" in the planning computed tomography (CT) images.

However, in contrast to OARs segmentation, the task of target volume delineation, which also needs to be contoured on the planning CT images but often requires consideration of clinical information beyond the visual features, remains crucial for treatment planning and has traditionally been the responsibility of experienced radiation oncologists. This task is perceived as more challenging due to its intrinsic need for the integration of multimodal knowledge. Although a multitude of segmentation models have been proposed and explored to enhance the precision and efficacy of this task over the last few years[4–6], a conspicuous gap in research persists, particularly regarding multimodal target delineation[3].

[1]Department of Radiology, Massachusetts General Hospital (MGH) and Harvard Medical School, Boston, MA, USA. [2]Department of Radiation Oncology, Yonsei University College of Medicine, Seoul, South Korea. [3]Institute for Innovation in Digital Healthcare, Yonsei University, Seoul, South Korea. [4]Department of Radiation Oncology, Yongin Severance Hospital, Yongin, Gyeonggi-do, South Korea. [5]Department of Radiation Oncology, Gangnam Severance Hospital, Seoul, South Korea. [6]Oncosoft Inc., Seoul, South Korea. [7]Kim Jaechul Graduate School of AI, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. [8]These authors contributed equally: Yujin Oh, Sangjoon Park. ✉e-mail: jinsung@yuhs.ac; jong.ye@kaist.ac.kr

This is because the delineation of radiation therapy target transcends beyond the mere consideration of visual elements, such as the gross tumor volume (GTV)[7], and necessitates the incorporation of a myriad of factors, including tumor stage, histological diagnosis, the extent of metastasis, and gene mutation. These factors critically influence the potential for occult metastases, which may compromise the survival outcome of a patient. Areas at elevated risk for such metastatic growth are often treated electively, necessitating clinical consideration that is deeply rooted in a comprehensive understanding of various data modalities. Furthermore, additional factors, such as a patient's performance status and age, which collectively contribute to the general condition, also exert an impact on treatment target delineation. Given the imperative nature of considering information beyond imaging in target volume delineation, the application of a multimodal approach in radiation oncology is not merely beneficial but essential for the tasks of the radiation oncology[8]. This is particularly substantiated by the necessity to incorporate textual clinical data, which can significantly influence the identification and subsequent treatment of regions susceptible to occult metastases.

Recently, large language models (LLMs)—AI models proficient in processing and generating text, code, and other data types—have witnessed remarkable advancements[9–11]. Trained on extensive datasets of text and code, these models discern relationships among varied data types and generate new data, adhering to learned patterns. Furthermore, multimodal data such as images, signals, etc., can be easily integrated into LLMs through adapters and generative models for vision understanding and generation, respectively. Consequently, these models have demonstrated promise in a myriad of medical tasks, including multimodal medical report generation, medical question answering, and multimodal segmentation with medical images like chest X-rays[12–15].

Inspired by the multimodal integration capability of LLMs and needs for multimodal information for tumor target delineation, here we present a 3-dimensional (3D) multimodal clinical target volume (CTV) delineation model, LLMSeg, by integrating clinical information through the LLM for conditioning a segmentation model. Specifically, by leveraging the textual information from well-trained LLMs through simple prompt tuning, our cross-attention-based segmentation model has adeptly integrated text-based clinical information into the target volume contouring task. More specifically, as illustrated in Fig. 1a, we introduce an interactive alignment framework which uses both self-attention and cross-attention mechanisms in a bidirectional manner (text-to-image and image-to-text features), by following the concept of promptable segmentation from Segment Anything Model (SAM)[16]. To further improve the quality of features, we implement this interactive alignment between all the skip-connected image encoder features with the LLM feature. These layer-wise multimodal features are then combined to jointly predict the target labels through the multimodal decoder. In this way, we ensure the image encoder to efficiently extract meaningful text-related representations and vice versa. Finally, to transfer the LLM's knowledge while the entire network parameters are kept and achieve superior performance in various downstream tasks[17–19], we adapt the idea of light-weight learnable text prompts to

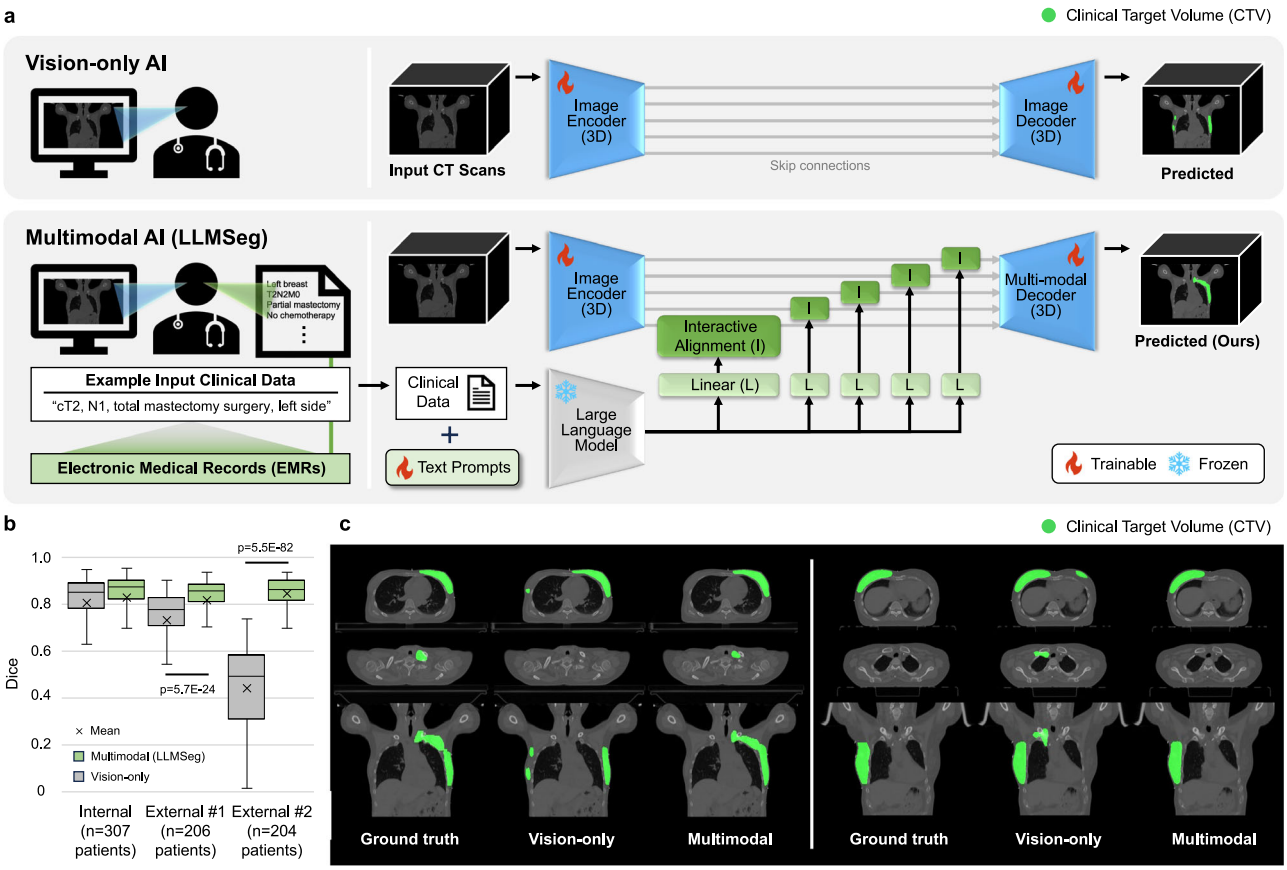

Fig. 1 | Overview of our proposed LLMSeg. a Illustration comparing the concept between the traditional vision-only AI and the multimodal AI in the context of radiotherapy target volume delineation. **b** Quantitative comparison of CTV contouring performance in the Dice metric. The Dice metric for each trial is presented with whiskers representing the range from minimum to maximum values. The center line indicates the median, the bounds of the box represent the interquartile range (from the lower quartile to the upper quartile), and the x mark indicates the mean. *n* denotes the number of patients. The *p* values indicate the statistically significant superiority of the proposed multimodal LLMSeg. All statistical tests were two-sided. **c** Visual assessment of each concept. Source data are provided as a Source Data file.

fully leverage the great linguistic capability of the LLM within the proposed multimodal AI framework.

In this work, we apply LLMSeg to the breast cancer target volume delineation task to evaluate its context-aware radiotherapy target delineation performance compared to a unimodal AI. Additionally, we expand its application to prostate cancer cases. By utilizing a well-curated, large-scale dataset from three institutions for development and external validation, we verify its capability to integrate pivotal clinical information, such as tumor stage, surgery type, and laterality. Experimental results confirm that the model not only demonstrates a significantly enhanced target contouring performance compared to existing unimodal segmentation models but also exhibits behavior that contours targets in accordance with provided clinical information. Notably, the model exhibits superior performance enhancement on an external dataset and shows stable performance gains in data-insufficient settings, demonstrating generalizability and data-efficiency that are not only apt for the characteristics of medical domain data but also aligns well with the perspective of clinical experts.

Results

Accurate and robust CTV delineation performance of multimodal model

Figure 1b presents a comparative analysis between the vision-only model and our proposed multimodal model for CTV delineation in breast cancer patients for all the validation sets. For internal validation, both methods showed promising performance of above 0.8 in the Dice metric, with a substantial improvement is observed in ours. However, the vision-only model showed a drastic performance drop of 0.73 and 0.44 in the Dice metric in both external settings. Specifically, in the case of external set #2, where the manufacturer of acquisition modality differs from that of internal and external set #1, the vision-only model completely failed to perform CTV delineation. Despite encountering visually shifted data distributions, our multimodal model demonstrated notable stability by consistently maintaining performance across all experimental conditions.

We qualitatively compare two different approaches in Fig. 1c. In general, CTV for breast cancer radiation therapy can be categorized into two primary types: one that involves treatment of the breast or chest wall alone, and the other that electively treats the regional lymph nodal area (including axillary, supraclavicular, and internal mammary lymph nodes (LNs)) in addition to the aforementioned areas, given the frequent metastasis of breast cancer to these regions. On the left side of Fig. 1c, despite the ground truth label posing CTV on both the breast and regional LNs, the vision-only model only contours the breast alone. Moreover, as the vision-only model lacks information about the laterality of the breast that diagnosed as cancer, partial segmentation masks are observed on the opposite breast. In contrast, the multimodal model accurately contours the breast and regional LNs that need to be treated as CTV. On the right side of Fig. 1c, despite early breast cancer case requiring treatment of the breast only, the vision-only model incorrectly includes the regional LNs as CTV. Moreover, CTVs are extended to the opposite breast. On the other hand, the multimodal model that integrates the clinical information accurately contours the requisite treatment areas, encompassing both the breast and the regional LNs, aligning with the ground truth.

We further compared our method with other diverse vision-only and multimodal methods in Table 1. Our proposed context-aware segmentation, in which the given textual information is not explicitly visible as an actual object in the input image, compared to traditional vision-language segmentation[20,21]. Therefore, we adapted publicly available 2D text-driven multimodal segmentation frameworks from various segmentation categories as our baseline models[22–24]. Furthermore, we conducted comparisons with two advanced visual backbones[25,26] to justify our selection of the 3D residual U-Net as the visual backbone. In the results shown in Table 1, HIPIE[22], and LISA[23], considered SOTA models for 2D referring and reasoning segmentation respectively, showed suboptimal performance in 3D context-aware segmentation. On the other hand, ConTEXTualNet[24], capable of handling 3D images as inputs, showed promising performance. Nevertheless, our approach demonstrated the SOTA performance across all evaluation metrics in various validation settings.

Performance evaluation by expert reveals superiority of multimodal model

The assessment of the target volume should not be based on mere metric evaluations such as the Dice, but rather by appropriate clinical

Table 1 | Comparison of 3D CTV delineation performance for breast cancer patients

		Vision-only AI			Multimodal AI			
Dataset	Metric	3D ResUNet[39]	3D SegMamba[26]	3D UNETR[25]	HIPIE[22]	LISA[23]	ConTEXTualNet[a24]	LLMSeg (Ours)
Internal test (*N* = 307)	Dice ↑	0.807	0.699	0.592	0.743	0.746	0.819	0.829
		(0.788–0.825)	(0.679–0.718)	(0.576–0.606)	(0.732–0.754)	(0.731–0.760)	(0.800–0.835)	(0.809–0.845)
	IoU ↑	0.698	0.559	0.433	0.600	0.608	0.715	0.730
		(0.677–0.718)	(0.538–0.580)	(0.418–0.447)	(0.587–0.613)	(0.591–0.624)	(0.695–0.733)	(0.709–0.748)
	HD-95 ↓	6.674	14.857	18.408	3.479	4.437	4.540	3.386
		(5.891–7.452)	(14.258–15.483)	(17.930–18.918)	(3.139–3.850)	(3.915–4.995)	(3.806–5.273)	(2.890–3.949)
External test #1 (*N* = 206)	Dice ↑	0.731	0.555	0.522	0.736	0.691	0.815	0.822
		(0.707–0.755)	(0.523–0.587)	(0.508–0.535)	(0.719–0.751)	(0.669–0.712)	(0.798–0.832)	(0.805–0.836)
	IoU ↑	0.599	0.422	0.359	0.594	0.547	0.701	0.709
		(0.575–0.622)	(0.392–0.451)	(0.347–0.370)	(0.576–0.611)	(0.524–0.569)	(0.680–0.721)	(0.689–0.727)
	HD-95 ↓	19.922	16.451	22.677	4.973	10.859	6.362	4.256
		(18.611–21.189)	(15.725–17.152)	(21.794–23.599)	(4.299–5.680)	(9.814–11.926)	(5.084–7.712)	(3.471–5.176)
External test #2 (*N* = 204)	Dice ↑	0.444	0.638	0.565	0.617	0.532	0.826	0.844
		(0.419–0.469)	(0.619–0.658)	(0.554–0.576)	(0.593–0.640)	(0.502–0.560)	(0.809–0.840)	(0.826–0.857)
	IoU ↑	0.302	0.484	0.399	0.469	0.389	0.715	0.740
		(0.282–0.322)	(0.464–0.507)	(0.388–0.409)	(0.446–0.492)	(0.362–0.414)	(0.697–0.732)	(0.722–0.756)
	HD-95 ↓	33.339	16.434	17.154	12.805	15.625	5.179	3.004
		(32.997–33.615)	(15.935–16.904)	(16.761–17.486)	(11.975–13.600)	(14.931–16.257)	(4.250–6.160)	(2.555–3.533)

[a]Modified for 3D CT segmentation.

Table 2 | Expert evaluation of CTV delineation performance for breast cancer patients

Dataset	Expert rubrics					
	Laterality (1 point)	Surgery type (1 point)	Volume definition (1.5 point)	Coverage (1 point)	Integrity (0.5 point)	Total (5 point)
	Vision-only AI					
Internal test (*N* = 307)	0.786	0.887	0.900	0.478	0.216	3.267
	(0.738–0.833)	(0.854–0.918)	(0.844–0.959)	(0.436–0.518)	(0.190–0.243)	(3.139–3.385)
External test #1 (*N* = 206)	0.344	0.863	0.680	0.186	0.124	2.198
	(0.279–0.412)	(0.821–0.899)	(0.615–0.748)	(0.149–0.226)	(0.093–0.154)	(2.056–2.346)
External test #2 (*N* = 204)	0.268	0.828	0.488	0.029	0.087	1.700
	(0.213–0.332)	(0.782–0.874)	(0.418–0.554)	(0.015–0.047)	(0.062–0.111)	(1.567–1.832)
	Multimodal AI (LLMSeg)					
Internal test (*N* = 307)	0.990	0.987	1.142	0.602	0.253	3.973
	(0.977–1.000)	(0.975–0.995)	(1.092–1.188)	(0.562–0.641)	(0.223–0.280)	(3.887–4.059)
External test #1 (*N* = 206)	0.990	0.983	1.174	0.532	0.260	3.939
	(0.976–1.000)	(0.963–0.998)	(1.105–1.243)	(0.480–0.581)	(0.226–0.294)	(3.821–4.059)
External test #2 (*N* = 204)	0.990	0.986	1.173	0.611	0.250	4.010
	(0.975–1.000)	(0.970–0.998)	(1.111–1.233)	(0.562–0.663)	(0.215–0.287)	(3.889–4.116)

rationale. In the context of breast contouring, this involves considerations such as whether the target volume has been contoured on the treated side of the breast, the contouring performed on the breast or chest wall depending on the type of surgery (breast-conserving surgery (BCS) or mastectomy), and whether the regional LNs have been included. Therefore, the appropriateness of target contouring should be evaluated by a board-certified radiation oncologist, ensuring a clinically relevant perspective in the assessment. To this end, five rubrics (laterality, surgery type, volume definition, coverage, integrity) were suggested by the board-certified radiation oncologists, to objectively and specifically evaluate the target volume with differentiated scoring reflecting their importance. Detailed descriptions of these rubrics are available in Supplementary Table 1 with Supplementary Fig. 1.

When evaluated using the proposed rubrics as indicated in Table 2, the multimodal model exhibited superior performance, achieving total scores up to twice as high as those of the vision-only model. Importantly, the model exhibited notably larger gains in rubrics like laterality and volume definition, where incorporation of the clinical context is crucial to achieve accurate results, than in metrics indicative of contouring quality, such as coverage and integrity. This performance gain was particularly pronounced in the external validation, notably in external set #2, where differences in the image acquisition setting were noted. This demonstrates the multimodal model's robustness and clinical relevance across varied datasets and potential diverse clinical scenarios.

Data efficiency and robustness of the multimodal model

During the training process of clinical specialists, learning is expedited when textual clinical information is integrated alongside imaging studies, as opposed to focusing on target volume in images alone. This approach facilitates a more rapid assimilation of tendencies and principles of target volume contouring, enabling effective learning even with fewer cases. We sought to determine whether this efficiency of learning through the integration of textual clinical information could be applied to our multimodal approach.

We observed the performance of each concept in target volume contouring by progressively reducing the size of training dataset. As illustrated in Fig. 2a, our multimodal model demonstrated its data efficiency by maintaining stable performance above 0.8 in the Dice even with 40% of data availability. This starkly contrasts with the vision-only model, whose performance dropped from initial Dice of 0.8–0.7. When utilizing only 20% of the training dataset, the multimodal model's performance decreased slightly below 0.8 in the Dice, while the vision-only model completely failed to contour CTV in the limited dataset scenario. This performance gap was particularly evident in external validation results. For external validation #1, the initial discrepancy between two models was ~0.1 in the Dice metric. However, as the size of training dataset decreased, the discrepancy doubled. For external validation #2, notable overfitting issues were observed in the vision-only model. On the contrary, our multimodal model achieved robust performance when trained with a reduced dataset of less than 40%. Qualitative analysis, as depicted in Fig. 2b, also supports these results. Detailed quantitative results for all metrics are further provided in Supplementary Table 2.

Differential target contouring based on varied textual inputs

To validate the hypothesis that our multimodal model genuinely performs CTV delineation based on textual clinical information, we conducted an experiment to assess whether altering the textual clinical information alone would yield different delineation results, even for the same CT, as illustrated in Fig. 3a.

As depicted in Fig. 3b, c, the model performed contouring different targets for the same CT, contingent on the provided clinical data. In Fig. 3b, for a patient with left breast cancer at stage T1N0M0, upstaging the T stage or N stage demonstrated the inclusion of regional LNs, and altering the tumor's laterality from left to right resulted in contouring on the opposite side. Interestingly, when the type of surgery was changed from BCS to total mastectomy, it was observed that the previously spared skin was no longer spared, and the target volume was expanded to include the chest wall. For another patient with right breast cancer at stage T2N1M0, as exemplified in Fig. 3c, downstaging N stage leads to the omission of regional LNs from the designated target volume, and changing the type of surgery to BCS results in a strategic shift to sparing the skin and excluding the pectoralis muscle from the treatment volume. These quantitative results align precisely with the decision policy of radiation oncologists, and substantiate that our model contours the target volume, strongly referencing the textual clinical information as well as the imaging features.

Exploring textual clinical information provision methods in the multimodal model

To demonstrate the necessity of LLM as for our textual clinical information provision method, we conducted an ablation study by

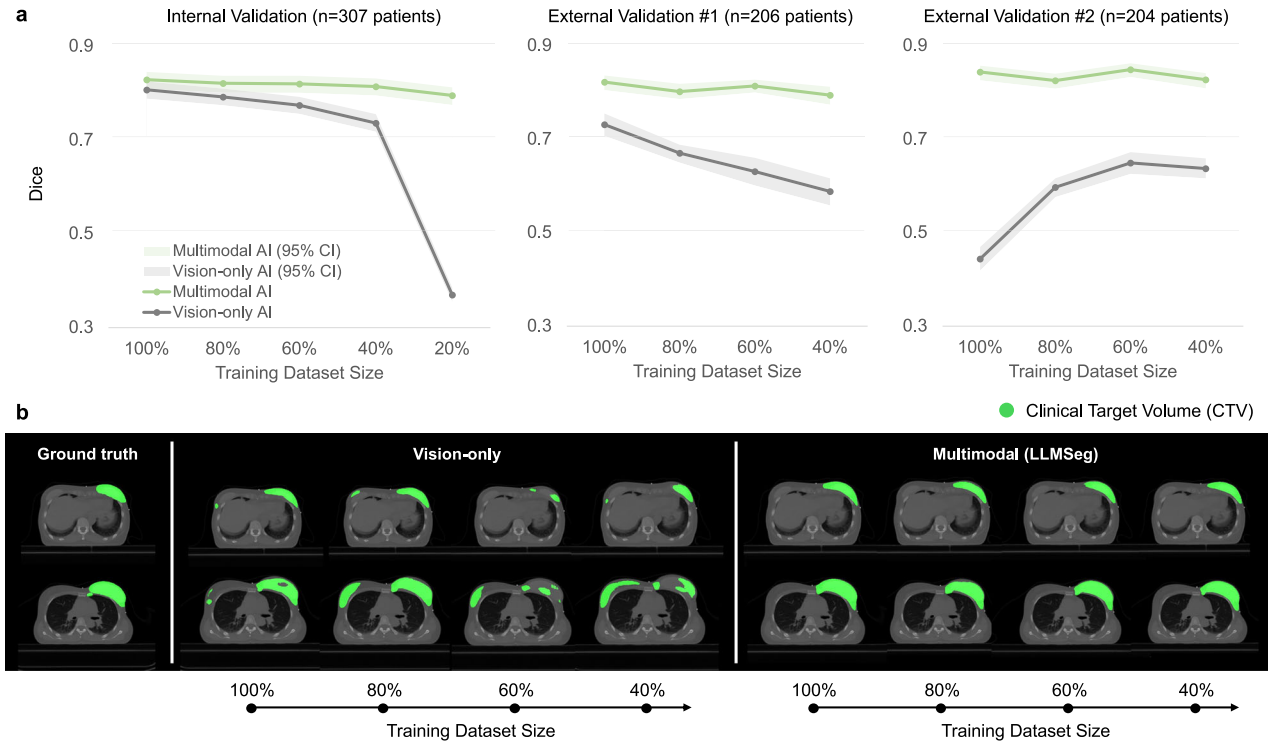

Fig. 2 | Comparison of target contouring performance based on varying training dataset sizes. a Quantitative comparison for all the validation sets. The Dice metric for each trial is presented as mean values (center lines) with 95th percentile of confidence intervals calculated with the non-parametric bootstrap method (shaded areas). *n* denotes the number of patients. **b** Visual comparison for external validation #1. Source data are provided as a Source Data file.

replacing our textual module by a simple numeric category method and a CLIP text encoder trained on a relatively smaller textual dataset compared to LLM[27]. As indicated in Table 3a, the numeric category method, by representing each clinical information as categorized numbers, exhibited promising performance and showed relatively marginal performance drops to our method in the internal validation setting. However, in the two external validations, the performance gaps were increased up to 0.1 in the Dice metric and significantly more in the HD-95 metric, of up to 10 cm. Moreover, when replacing the textual module with the CLIP ViT-B/16 while maintaining our proposed multiple text prompt tuning method, huge performance gaps were observed compared to our method of up to 0.3 in the Dice metric and up to 10 cm in the HD-95 metric. These findings indicate that the effectiveness of the proposed multimodal model originates from leveraging LLM.

Specifically, the numeric category method exhibited the second-most promising performance and showed relatively marginal performance drops to our method in the internal validation setting. However, in the two external validation settings, the performance gaps were increased. Hence, we qualitatively evaluated the source of the performance gap in Fig. 4a. In Case #1, where a patient underwent total mastectomy for T2N1M0 cancer in the left breast, our method accurately contoured the surgically treated breast with an implant, including the regional nodal area in the target volume. However, the numeric category method generated segmentation masks for both breasts, with more mask generation observed on the opposite breast. Similarly, in Case #2, where a patient underwent breast conservation surgery for T2N1M0 cancer in the left breast, our method accurately included the breast and regional nodes in the target volume while sparing the skin and chest wall. In contrast, the numeric category method only included the breast area in the target volume, excluding the regional nodes, and included parts of the skin and chest wall similar to the mastectomy case. Moreover, it partially generated segmentation masks on the opposite breast, demonstrating incomplete reflection of the clinical context.

We further ablated our employment of introducing clinical data by replacing it with various methodologies. These include utilizing a single or multiple text prompts through prompt tuning, low-rank adaptation (LoRA) fine-tuning[28], and directly employing a pre-trained LLM without tuning. As indicated in Table 3b, our proposed text prompt tuning method consistently outperformed those using LoRA fine-tuning and a no-tuning strategy. Moreover, employing multiple learnable text prompts showed an improved performance compared to using a single text prompt. These results indicate that the introduced learnable text prompts were optimized to efficiently fine-tune the LLM for the target volume contouring task.

Ablation study of input clinical data components

We further conducted ablation study by omitting each piece of input clinical information and compared the difference between a competing method (Numeric Category) and our method (LLMSeg) in Fig. 4b, c. Firstly, without omission as shown in Fig. 4b, our method accurately segmented only the right breast area as the target volume for a person with T1aN0M0 cancer who underwent BCS. However, when the information for T stage was removed, the model included some regional nodes in the target range, similar to cases with higher stages. This trend was similarly observed in the omission of N stage information, where the model included regional nodes as in cases with nodal metastasis like N1 or N2. Likewise, without information about laterality, the model inaccurately contoured the opposite breast. On the contrary, the competing model showed inaccurate results such as contouring on the opposite breast even without omission. Moreover, regardless of the presence or absence of omission, there was little change in target contouring (e.g., laterality), or target contouring changed in patterns unrelated to the omitted information (e.g., contouring on the opposite side when omitting T stage or N stage

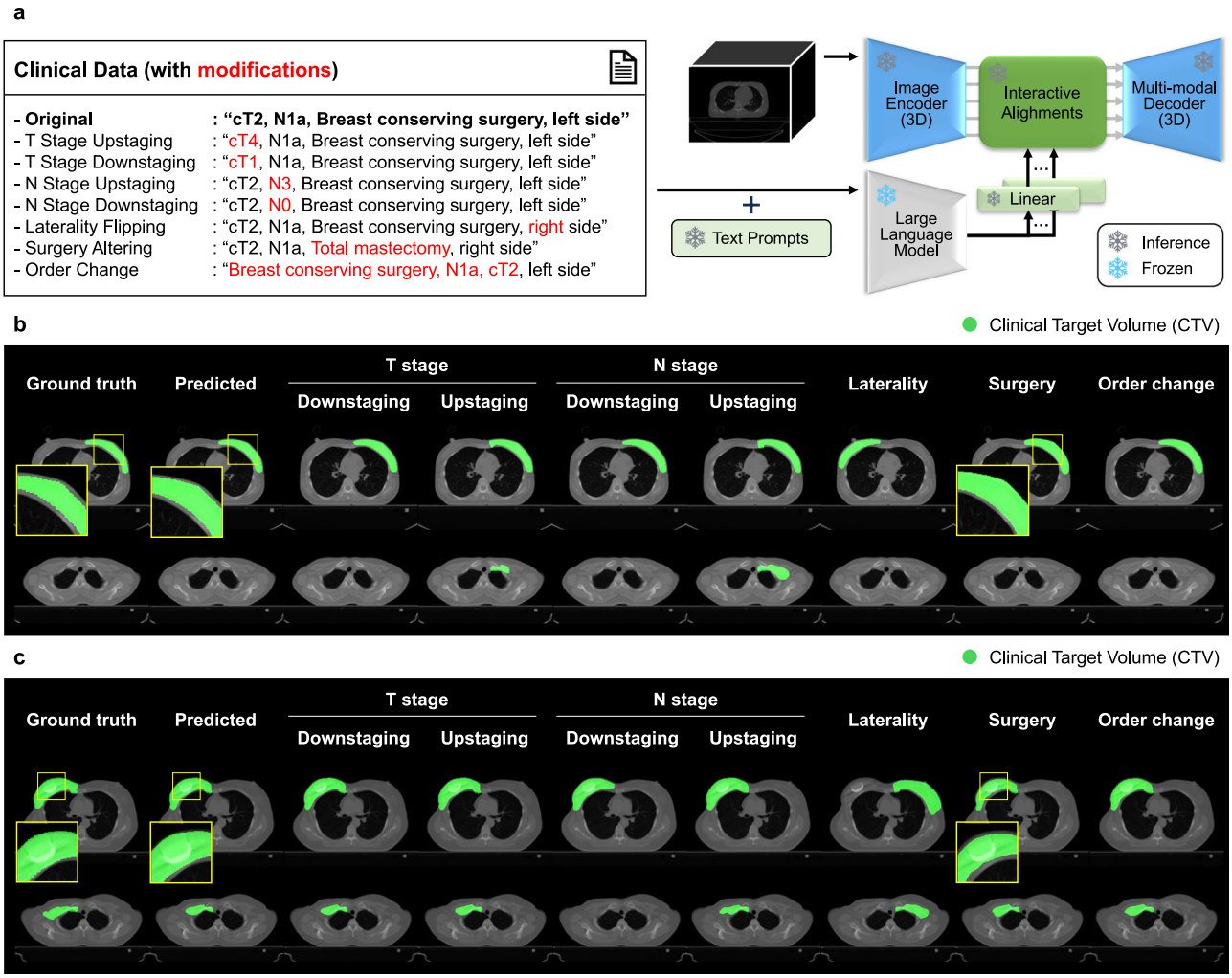

Fig. 3 | Analysis of clinical data alignment for target contouring. a Illustration of modification of the input clinical data, given the same CT scan. Red font indicates modified input text. **b**, **c** Visual assessment of radiotherapy target contouring with modified input clinical data.

information). These results indicate that the competing model receiving clinical context in a simpler manner failed to effectively incorporate such information and perform CTV delineation unrelated to the provided information. Similarly, in another case of T1cN1M0 breast cancer in the left breast where total mastectomy was performed as shown in Fig. 4c when surgery information is not provided, our method misidentified the surgery type and produced segmentation results resembling BCS, sparing the skin and chest wall. However, the competing model rather contoured on the opposite breast, which was irrelevant to surgical method.

In Table 4, we further assessed these ablation results quantitatively. For our method, the exclusion of information regarding laterality, which influences the decision on which breast to contour, resulted in the most significant decrease in performance. This was followed by similar degrees of performance decrease upon excluding information related to surgery type and N stage, which impact the inclusion of the skin, chest wall, and the regional nodes. Although excluding T stage information did result in a decrease in performance, it was the least significant, which is rational considering the minimal impact of T stage information on target volume delineation.

Overall, these comparative results suggest that our model considers the clinical context provided in text and is hindered in accurate target volume delineation if any component is missing. That is said, excluding any one component results in lower performance compared to using all available information, indicating that every component contributes to the model's performance.

Exploring other cancer types

We further evaluated the proposed multimodal target volume contouring for prostate cancer patients. For prostate cancer, clinical data were directly curated from EMR, as detailed in Supplementary Table 3. This curated EMR data, along with each patient's age, were then summarized as input clinical data. Similar to the breast cancer study, we observed the superiority of our multimodal approach over the vision-only approach, with a notable performance gain of up to 0.05 in Dice metric through all the validation settings as shown in Table 5.

Similar to breast cancer, an expert evaluation was conducted for prostate cancer. A rubric-based analysis of expert evaluation in Table 6 clearly showed effectiveness of our method. Particularly, these benefits became unequivocally evident in the external validation setting, showing more than double the differences in total scores. Among those necessitating in-depth reference to clinical information for precise scoring—notably, the delineation of the primary site (assessing prostate volume coverage, including the seminal vesicle) and the volume definition (evaluating regional node irradiation appropriateness)—exhibited significantly larger differences compared to the vision-only model. Details on the rubrics used for prostate cancer can be found in the Supplementary Fig. 2 and Supplementary Table 4.

https://doi.org/10.1038/s41467-024-53387-y

Table 3 | Ablation studies on network components

Components	Metric	LLMSeg (Ours)	(a) Textual module		(b) Tuning method ablation		
Text module		LLaMA-7B-chat[10]	Numeric category[a]	CLIP ViT-B/16[27]	LLaMA-7B-chat[10]		
Tuning method		Text prompts	–	Text prompts	One text prompt	LoRA[28]	No tuning
Internal Test (*N* = 307)	Dice ↑	0.829	0.821	0.813	0.822	0.829	0.819
		(0.809–0.845)	(0.805–0.836)	(0.795–0.830)	(0.803–0.839)	(0.812–0.844)	(0.799–0.835)
	IoU ↑	0.730	0.715	0.706	0.721	0.726	0.715
		(0.709–0.748)	(0.697–0.733)	(0.687–0.725)	(0.700–0.739)	(0.708–0.742)	(0.695–0.732)
	HD-95 ↓	3.386	5.387	5.769	4.036	3.923	4.546
		(2.890–3.949)	(4.699–6.129)	(5.032–6.500)	(3.442–4.697)	(3.352–4.543)	(3.877–5.210)
External Test #1 (*N* = 206)	Dice ↑	0.822	0.734	0.737	0.817	0.809	0.825
		(0.805–0.836)	(0.710–0.755)	(0.710–0.761)	(0.799–0.832)	(0.793–0.825)	(0.812–0.838)
	IoU ↑	0.709	0.601	0.609	0.703	0.692	0.712
		(0.689–0.727)	(0.577–0.623)	(0.583–0.632)	(0.683–0.722)	(0.673–0.711)	(0.695–0.729)
	HD-95 ↓	4.256	17.010	15.484	4.914	7.838	9.148,
		(3.471–5.176)	(15.829–18.165)	(13.998–17.016)	(3.908–5.968)	(6.704–8.895)	(7.812–10.569),
External Test #2 (*N* = 204)	Dice ↑	0.844	0.737	0.528	0.832	0.796	0.829
		(0.826–0.857)	(0.717–0.755)	(0.490–0.567)	(0.815–0.846)	(0.779–0.810)	(0.812–0.842)
	IoU ↑	0.740	0.599	0.409	0.724	0.674	0.718
		(0.722–0.756)	(0.577–0.618)	(0.375–0.444)	(0.705–0.739)	(0.654–0.690)	(0.700–0.733)
	HD-95 ↓	3.004	13.703	15.429	3.056	9.650	5.078
		(2.555–3.533)	(12.706–14.724)	(14.312–16.583)	(2.673–3.507)	(8.683–10.623)	(4.416–5.827)

[a]Numeric categorization of text, e.g., "0301", corresponds to "N0" for the first digit, "T3" for the second, "mastectomy" for third, and "right" for the last.

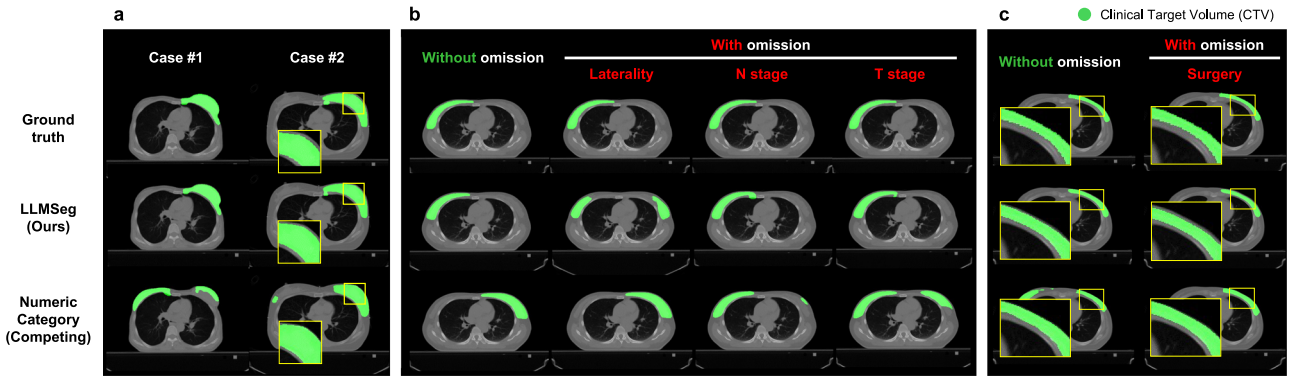

Fig. 4 | Qualitative comparison of different multimodal methods with omitted clinical data components. a Comparison with numeric category method: Case 1 (left breast, T2N1M0, post-mastectomy) and Case 2 (left breast, T2N1M0, post-breast conservation surgery) show our method (LLMSeg) accurately includes surgically treated areas and regional nodes, while the numeric category method inaccurately segments both breasts, missing clinical context. **b** Omission experiment for tumor information: For right breast T1aN0M0 cancer, our method segments accurately without omission. Omitting T stage, N stage, or laterality causes incorrect regional node inclusion or opposite breast contours. The competing method is inaccurate regardless of omission. **c** Omission experiment for surgery information: In left breast T1cN1M0 cancer post-mastectomy, our method without surgery information mimics breast-conserving surgery. The competing method inaccurately contours the opposite breast irrespective of surgery information.

Discussion

Despite the promising outcomes demonstrated by AI models in various studies, a notable limitation prevalent in the field of medical AI has been the predominant development of models tailored for singular, specialized tasks[29]. For instance, models have been specifically designed and trained to excel in a singular task, such as segmentation[4,6], diagnosis[30,31], or prognosis prediction[32,33], without the adaptability to transition across various tasks. While these specialized models perform commendably within their designated task, they lack the flexibility to navigate the complex challenges in the medical domain, where the ability to integrate, and concurrently process diverse tasks is crucial.

In the nascent stages of applying vision-language models to the medical domain, initial research endeavors have predominantly focused on the most simple form of vision-text paired data, such as chest radiographs[34]. These studies have explored various tasks, including zero-shot classification[35], report generation[36,37], and text-guided segmentation[15,24]. However, the field of radiation oncology emerges as a particularly potent application area for such models[8]. Radiation oncology exemplifies a robust case for the adoption of multimodality, underpinned by two fundamental factors[1]. Firstly, decision-making in Radiation Oncology, especially in determining treatment scope and dose, extends beyond imaging to include a plethora of clinical information, such as surgical notes, pathology reports, and electronic medical records, which can be conveyed textually. Secondly, the integration of prior knowledge, including standard treatment guidelines and radiation oncology textbooks, is vital for informed

Table 4 | Ablation of input clinical data components for two different multimodal methods

	Ablated input text components				
Dataset	**No ablation**	**Laterality**	**N stage**	**T stage**	**Surgery**
	LLMSeg (Ours)				
Internal test (*N* = 307)	0.829	0.424	0.799	0.816	0.791
	(0.809–0.845)	(0.383–0.467)	(0.781–0.815)	(0.795–0.832)	(0.771–0.809)
	0.730	0.345	0.686	0.712	0.680
	(0.709–0.748)	(0.309–0.382)	(0.667–0.702)	(0.690–0.730)	(0.659–0.699)
	3.386	12.751	5.071	3.934	5.050
	(2.890–3.949)	(11.894–13.530)	(4.570–5.646)	(3.415–4.522)	(4.531–5.583)
External test #1 (*N* = 206)	0.822	0.344	0.786	0.804	0.710
	(0.805–0.836)	(0.305–0.385)	(0.773–0.798)	(0.787–0.820)	(0.677–0.740)
	0.709	0.251	0.656	0.686	0.590
	(0.689–0.727)	(0.219–0.283)	(0.641–0.671)	(0.666–0.705)	(0.559–0.619)
	4.256	14.093	7.424	6.981	8.652
	(3.471–5.176)	(13.197–14.991)	(6.450–8.481)	(5.946–8.065)	(7.724–9.616)
External test #2 (*N* = 204)	0.844	0.390	0.805	0.833	0.763
	(0.826–0.857)	(0.340–0.434)	(0.790–0.817)	(0.816–0.846)	(0.740–0.784)
	0.740	0.304	0.682	0.724	0.637
	(0.722–0.756)	(0.263–0.341)	(0.667–0.696)	(0.706–0.739)	(0.613–0.660)
	3.004	11.971	4.618	3.954	7.246
	(2.555–3.533)	(11.161–12.814)	(4.244–5.029)	(3.475–4.475)	(6.571–7.951)
	Numeric category[a] (Competing method)				
Internal test (*N* = 307)	0.821	0.720	0.785	0.793	0.791
	(0.805–0.836)	(0.695–0.744)	(0.768–0.800)	(0.777–0.808)	(0.771–0.809)
	0.715	0.599	0.664	0.674	0.680
	(0.697–0.733)	(0.573–0.624)	(0.646–0.680)	(0.657–0.691)	(0.659–0.699)
	5.387	8.758	6.042	5.932	5.050
	(4.699–6.129)	(7.874–9.715)	(5.346–6.802)	(5.198–6.707)	(4.531–5.583)
External test #1 (*N* = 206)	0.734	0.712	0.724	0.716	0.710
	(0.710–0.755)	(0.690–0.734)	(0.695–0.749)	(0.694–0.738)	(0.677–0.740)
	0.601	0.573	0.595	0.579	0.590
	(0.577–0.623)	(0.551–0.596)	(0.567–0.620)	(0.557–0.603)	(0.559–0.619)
	17.010	19.446	18.058	20.579	8.652
	(15.829–18.165)	(18.102–20.765)	(16.906–19.214)	(19.284–21.768)	(7.724–9.616)
External test #2 (*N* = 204)	0.737	0.555	0.668	0.669	0.700
	(0.717–0.755)	(0.513–0.595)	(0.638–0.694)	(0.638–0.698)	(0.681–0.718)
	0.599	0.436	0.529	0.538	0.553
	(0.577–0.618)	(0.401–0.469)	(0.502–0.555)	(0.507–0.567)	(0.534–0.573)
	13.703	13.026	11.460	13.082	11.504
	(12.706–14.724)	(12.144–13.812)	(10.562–12.343)	(11.945–14.242)	(10.538–12.383)

[a]For numeric categorization method, each digit is ablated, e.g., "0301", where the first 0 representing "N0" is replaced with a character "?" to yield "?301".

Table 5 | Comparison of 3D CTV delineation performance for prostate cancer patients

Metric	Vision-only AI			Multimodal AI (LLMSeg)		
	Dice ↑	**IoU ↑**	**HD-95 ↓**	**Dice ↑**	**IoU ↑**	**HD-95 ↓**
Internal test (*N* = 189)	0.725	0.598	3.522	0.754	0.631	3.036
	(0.697–0.751)	(0.568–0.626)	(3.084–4.014)	(0.729–0.779)	(0.604–0.658)	(2.646–3.482)
External test #1 (*N* = 141)	0.682	0.535	3.762	0.729	0.589	3.522
	(0.656–0.705)	(0.508–0.560)	(3.463–4.102)	(0.706–0.750)	(0.563–0.613)	(3.194–3.932)

treatment decision-making, with these guidelines also being expressible in textual formats. Consequently, the necessity for multimodality is markedly emphasized in Radiation Oncology (see Supplementary Fig. 3).

Consequently, we have applied LLMs in our research. Our model introduces several aspects with substantial clinical value and has demonstrated commendable results by accurately segmenting radiation therapy target volume based on clinical information, thereby

Table 6 | Expert evaluation of CTV delineation performance for prostate cancer patients

Dataset	Expert rubrics				
	Primary site (1 point)	Volume definition (1.5 point)	Coverage (1 point)	Integrity (0.5 point)	Total (4 point)
	Vision-only AI				
Internal test (*N* = 189)	0.470	0.717	0.313	0.171	1.670
	(0.412–0.529)	(0.644–0.783)	(0.262–0.361)	(0.136–0.206)	(1.527–1.810)
External test #1 (*N* = 141)	0.266	0.424	0.115	0.090	0.895
	(0.212–0.320)	(0.367–0.482)	(0.079–0.147)	(0.061–0.122)	(0.781–1.007)
	Multimodal AI (LLMSeg)				
Internal test (*N* = 189)	0.583	0.951	0.379	0.249	2.162
	(0.529–0.639)	(0.874–1.027)	(0.326–0.428)	(0.211–0.283)	(2.003–2.310)
External test #1 (*N* = 141)	0.578	0.889	0.248	0.209	1.923
	(0.507–0.640)	(0.791–0.978)	(0.205–0.295)	(0.169–0.248)	(1.752–2.097)

achieving absolute performance where the multimodal model surpasses the vision-only model. It also exhibits a pronounced performance differential in external validation settings and demonstrates data efficiency in data-insufficient settings. This resonates intriguingly with the clinical implications, especially mirroring the learning trajectory and characteristics of clinical experts. In the clinical training of experts, reliance is placed on multimodality information; learning is not confined to either images or text but is rather a confluence of both, facilitating the inference of text-image relationships and enabling effective learning even with relatively fewer cases. This aspect of the clinical learning paradigm, being data-efficient, aligns seamlessly with our proposed multimodal model.

The decrement in classical AI-driven delineation generalization performance is often attributed to variations in image acquisition settings and characteristics of devices from different vendors, among other factors. Nonetheless, the ability of clinical experts to perform target contouring is scarcely influenced by external factors such as CT scanning conditions. This is because linguistical concepts embodied in textual clinical information, are independent of such acquisition settings. Therefore, it is plausible that our model, which learns in conjunction with such textual clinical information by leveraging the great linguistic capability of LLMs, demonstrates particularly commendable performance in external validation settings. This characteristic is particularly optimal for the medical domain, where training data is often limited and stable generalization performance is a prerequisite across varied external settings, thereby heralding a promising future for the application of multimodal models in medical AI.

Furthermore, we have demonstrated the necessity of incorporating clinical information into target volume contouring, particularly in cases such as breast cancer where the GTV may not be clearly visible in the planning CT image. This necessity is highlighted through diverse comprehensive qualitative comparison. In Fig. 1c, where the inclusion of clinical context is crucial for both cases, the multimodal target contouring reflects comprehensive considerations of clinical context. This necessity becomes more evident, where the absence of clinical context in the vision-only model results in clear failure cases. Additionally, in the detailed rubric comparison between the vision-only model and our multimodal model presented in Tables 2 and 6, the largest gains are observed in metrics that can be achieved through clinical considerations, such as laterality and volume definition. These results further emphasize the value of our multimodal approach.

Our study has several limitations. First, our evaluation is confined to patients at their initial diagnosis, leaving a scope for further exploration into varied patient scenarios and treatment stages, which can potentially influence the model's applicability and performance. Second, the model does not incorporate considerations for radiation therapy doses in target volume contouring, presenting an opportunity to explore how dose-related variables could be integrated to enhance delineation and treatment planning in future studies. Third, while the model utilizes refined, rather than raw, clinical data, future research can explore mechanisms for automating the data refinement process or further develop capabilities to process raw clinical data, thereby reducing the need for manual intervention and potentially uncovering additional insights from unstructured clinical reports. Fourth, although our research scope covers both breast and prostate cancers to confirm its applicability in various cancer types, these cancer types are categorized as having relatively standardized target volume. This suggests the necessity for further validation of our method's generalizability across a wider range of cancer types, which demand more challenging and intricate clinical considerations for accurate target volume delineation. Fifth, in our work, we focus CTV contouring to clearly demonstrate advantages of our multimodal model. However, GTV delineation, which involves contouring visually apparent areas, is crucial in clinical practice due to its importance in boost techniques for increased dose administration in many cancer types.

Additionally, in cancer types where the target volume is primarily determined based on GTV, such as lung cancer[38], the benefits of integrating clinical information through our method may be relatively limited. Therefore, it is necessary to validate whether our method still offers utility in such cancer types, where the emphasis is on GTV for target volume definition. Therefore, future studies should expand to encompass GTV contouring, thereby improving its clinical utility. Last, but not least, the black-box nature of AI may hinder clinician's direct utilization. Therefore, our proposed model should provide explainable results such as a confidence map in the clinical practice, as shown in Supplementary Fig. 4. These visual clues enable clinicians to interpret the model output by referencing the level of confidence for each segment of contour.

Despite aforementioned limitations, our research serves as a pivotal step towards the multimodal models in the field of radiation oncology, verifying the clinical utility and emphasizing the significance of intertwining textual clinical data with medical imaging. The model proposes a pathway for crafting more adaptable and clinically pertinent AI models in medical imaging and treatment planning. Future research would refine and broaden such models, closer to harnessing the full potential of multimodal framework in elevating clinical decision-making and patient care.

Methods

Ethic committee approval

The hospital data deliberately collected for this study were ethically approved by the Institutional Review Board of Department of Radiation Oncology at Yonsei Cancer Center, Department of Radiation Oncology at Yongin Severance Hospital, and Department of Radiation

Oncology at Gangnam Severance Hospital (approval numbers of 4-2023-0179, 9-2023-0161 and 3-2023-0396 for each). The requirement for informed consent was waived due to the retrospective nature of the study.

Schematic comparison of the workflows of radiology and radiation oncology

Supplementary Fig. 3 delineates the clinical workflows in Radiology and Radiation Oncology. In radiology, while the patient's history, previous diagnoses, past treatments, and previous imaging results are comprehensively considered, the most crucial element remains the findings visible in the current images, thus heavily relying on the visual information of the current imaging study. Conversely, in radiation oncology, determining the treatment target volume and prescribing doses necessitates a more comprehensive consideration of the patient's history, pre-and post-operative imaging results, surgical pathology findings, laboratory results, and other clinical information, resulting in a relatively less reliance on the current simulation CT images.

Additionally, the integration of prior knowledge, including standard treatment guidelines and radiation oncology textbooks, is crucial for informed treatment decision-making and can also be expressed in textual formats. Therefore, the significance of multimodal approach is notably enhanced in Radiation Oncology compared to Radiology.

Definition of task

In radiation oncology, the treatment target volumes are categorized into GTV, CTV, and Planning Target Volume (PTV). GTV corresponds to the visible tumor, aligns with traditional segmentation's objective to delineate visible image portions. CTV, while occasionally derived directly from GTV in the presence of a gross tumor, often also includes regions prone to microscopic disease. This necessitates the incorporation of diverse clinical factors, such as tumor type, histological findings, cancer stage (TNM classification), patient age, and performance status in specific cases. PTV further expands upon CTV to include margins that account for uncertainties in patient setup and positioning. Consequently, achieving accurate target volume delineation in radiation oncology goes beyond the scope of traditional segmentation tasks, necessitating incorporation of various clinical contexts as well as the structures visible on the CT scan.

Taking breast cancer as an example, in early-stage cases (e.g., stage I) where there is no regional LNs metastasis, often only the whole breast is included in the radiation therapy target volume. On the other hand, in advanced stages (e.g., stage IIIB), where regional LN metastasis is identified during surgery, there is often a need for elective nodal irradiation across all regional nodal areas. However, such distinctions are not discernible during the CT simulation for post-operative radiation therapy planning and require acquisition through other forms of information. Consequently, we aimed to develop a model that can consider clinical information such as primary tumor type, stage, age, and performance status in a manner akin to an experienced radiation oncologist by providing such data in the form of textual information to a multimodal model.

Among the primary cancer types, we initially targeted breast cancer. This was predicated on the fact that breast cancer presents with relatively uniform guidelines for target delineation according to the clinical information including primary tumor location, size, and the presence of nodal metastasis, etc. Furthermore, the inter-observer variability in target delineation for breast cancer is also expected to be small compared with other cancer types. Within the task of radiation therapy target delineation for breast cancer, we exclusively incorporated cases of patients at their initial diagnosis of breast cancer. This decision was based on the understanding that treatments with aims such as salvage or palliative often exhibit significant variability according to the preferences of the physicians as well as the patients, and other circumstances.

Details of clinical target volume

For breast cancer, the CTV for early breast cancer (Tis-T2) without nodal metastasis at initial diagnosis is limited to the whole breast. For those with nodal metastasis or in cases of locally advanced breast cancer (T3-4), as well as T2 cases with adverse features without proper axillary dissection, regional node irradiation was primarily considered. The delineation of regional nodes, especially the level of inclusion for the supraclavicular lymph node, is defined according to the Radiation Therapy Oncology Group guidelines for cases identified with N2 or more nodal metastasis, and by the European Society for Radiotherapy and Oncology guidelines for instances with N1 or less nodal involvement.

For prostate cancer, the definition of the CTV involved a more complex consideration of factors. In the presence of pelvic LN, regional node irradiation was performed in conjunction with prostate bed radiation. The decision to perform elective nodal irradiation on the pelvic LNs was based on the National Comprehensive Cancer Network risk groups, taking into account a combination of factors such as T stage, Prostate-Specific Antigen (PSA) levels, and Gleason score, particularly for those classified within the very high and high-risk groups. However, in individuals aged 80 and over, consideration of age led to the omission of pelvic LN irradiation. In cases where pathologic or imaging findings confirmed seminal vesicle invasion, contouring was performed to include the prostate and extend to the seminal vesicles within the CTV.

Details of datasets

For model development and internal validation, we acquired data from 981 patients treated at the Department of Radiation Oncology at Yonsei Cancer Center between September 2021 and October 2023. These patients had been initially diagnosed with breast cancer and underwent radiation therapy post-curative surgery with the primary objective of preventing recurrence. To better reflect real clinical application, the ideal approach for external validation needs the use of patient data acquired under different conditions and with equipment from a different vendor. Therefore, we utilized data from 206 patients treated at the Department of Radiation Oncology at Yongin Severance Hospital. We further utilized data from 204 patients treated at the Department of Radiation Oncology at Gangnam Severance Hospital. We confirmed that the external cohort was non-overlapping with those included in the model development nor internal validation.

Supplementary Table 5 presents the characteristics of breast cancer patients for each dataset. Across the train, internal, and external validation sets, distributions of factors such as location and T stage were observed to be consistent. The proportion of patients with LN metastasis and those undergoing total mastectomy was higher in the train and internal validation sets than the external validation set. Furthermore, due to the more advanced stages of disease, the proportion of patients who underwent neoadjuvant chemotherapy prior to surgery was observed to be higher in the train and internal validation sets compared to the external validation sets. Consequently, the percentage of patients receiving irradiation to the chest wall and regional LNs was also higher in the train and internal validation sets compared to the external validation set. When compared to the training and internal validation sets, external set #1 exhibited similar imaging equipment and conditions. However, external set #2 presented differences in image acquisition conditions, such as vendor, filter type, and slice thickness.

For evaluating the proposed method for other cancer types, we further acquired data from 943 prostate cancer patients from Yonsei Cancer Center and 141 prostate cancer patients from Yongin Severance Hospital. We confirmed that the external cohort was non-overlapping with those included in the model development nor internal validation. Supplementary Table 6 presents characteristics of prostate cancer patients for each dataset. In terms of the distribution of T and N stages,

as well as Gleason scores, the training, internal validation, and external validation sets demonstrated a relatively uniform distribution. However, the initial PSA levels were found to be higher in the training and internal validation sets compared to the external validation set. Additionally, the proportion of individuals undergoing prostatectomy was also higher in the training and internal validation sets, which consequently led to a higher percentage of patients receiving radiotherapy with a definitive aim in the external validation set, while those in the training and internal validation sets were more likely to receive radiotherapy with a salvage aim. There were no significant differences in image acquisition settings across the datasets.

We not only utilized patient's simulation CT images and CTVs for radiation therapy, but also incorporated text-based clinical information that is essential for precise target delineation. This additional information included the location of the primary cancer, type of surgery undertaken, disease stage, and the status of nodal metastasis. The input clinical data was prepared by the tabular format derived from raw clinical data for breast cancer, as shown in Supplementary Table 3a. The resulting clinical context was then curated using custom criteria. Initially, these criteria were devised by a board-certified radiation oncologist. Subsequent refinement was achieved through ablation studies on the components to construct the most effective clinical information, and the resulting examples of input texts are illustrated in the right-most column.

Compared to breast cancer, by utilizing tabular structure of clinical data which is curated by clinicians, for prostate cancer, we directly curated input clinical information from EMR data, by utilizing 10-shot in-context learning strategy with a pre-trained LLM, as shown in Supplementary Table 3b. Then, the curated EMR Data and each patient's age were summarized as input clinical data in the right-most column. In the future study, a similar in-context learning approach can be applied to the breast cancer study for an automated framework.

Details of implementation

The schematic of our multimodal AI is illustrated in Fig. 1. For the image encoder/decoder and the LLM, we employed the 3D Residual U-Net[39] and the pre-trained Llama2-7B-chat[10] model, respectively. For the interactive alignment modules, we utilized the two-way transformer modules of SAM[16]. We further propose detailed multimodal AI framework as illustrated in Supplementary Fig. 5. We introduce three key components: (a) text prompt tuning, (b) multimodal interactive alignment, and (c) CTV delineation.

(a) Text prompt tuning

To efficiently fine-tune the LLM, we introduce N-text prompts $\mathcal{V} = \{v^n|_{n=1}^N\}$ as illustrated in Supplementary Fig. 5a, where each $v^n \in \mathbb{R}^{M\times D}$ consists of M vectors with the dimension D, which is same embedding dimension as the LLM. These learnable vectors are randomly initialized, and then consistently prepended to each of tokenized clinical data, which denoted as [TEXT] tokens. We additionally append a token, denoted as [SEG], which is intended to attend to all the aforementioned vectors and tokens. Here, the final prompted text input t can be formulated as follows:

$$t = \{v_1^n, v_2^n, \ldots, v_M^n, [\,\text{TEXT}\,], [\,\text{SEG}\,]\}. \tag{1}$$

Then, using the prompted text input t, the frozen LLM results the context embeddings $g \in \mathbb{R}^{N\times D}$ as output embeddings as for the inputted [SEG] token.

(b) Multimodal interactive alignment

To align the context embeddings g with the image embeddings $f_l \in \mathbb{R}^{H_l W_l S_l \times C_l}$, where f_l is the lth layer output of the 3D image encoder, H_l, W_l, and S_l correspond to height, width, and slice of the image embeddings, and C_l is the intermediate channel dimension of each lth layer output, we first project g to have the identical dimension with that of each f_l through layer-wise linear layer. As illustrated in Supplementary Fig. 5b, the linearly projected context embeddings $\bar{g}_l$ are then self-attended and crossly-attended with the image embedding f_l to result context-aligned image embeddings f_l^*. Detailed specifications of each lth layer embeddings and the interactive alignment module are listed in Supplementary Table 7.

(c) CTV delineation

After the multimodal interactive alignment, the context-aligned image embeddings f_l^* become inputs for the 3D image decoder. As illustrated in Supplementary Fig. 5c, for the final predicted output $\hat{y}$, we calculated the combination of the Cross-entropy (CE) loss and the Dice coefficient (Dice) loss by following:

$$\begin{aligned} &\min_{\mathcal{D},\mathcal{V}} \mathcal{L} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}}(\hat{y}, y) + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}(\hat{y}, y), \\ &\text{where } \mathcal{L}(\hat{y}, y) = -\mathbb{E}_{x\sim P_X}\left[y_i \log p(\hat{y}_i)\right], \end{aligned} \tag{2}$$

where $\mathcal{D}$ denotes our prospoed LLMSeg, $\mathcal{V}$ denotes multiple text prompts, $\lambda_{\text{ce},}$ and λ_{dice} are hyper-parameters for each CE loss and Dice loss, respectively. $y \in \mathbb{R}^{B\times HWS}$ is the 3D ground-truth CTV mask, where B denotes batch size, H, W, and S correspond to height, width, and slice of ground-truth CTV mask. $p(\hat{y}_i)$ denotes softmax probability of the ith pixel within the final predicted output $\hat{y} \in \mathbb{R}^{B\times HWS}$, which is defined as:

$$\hat{y} = \mathcal{D}(x, t) \tag{3}$$

where $x \in \mathbb{R}^{B\times HWS}$ is input 3D CT scan, t is prompted clinical data corresponds to the input CT scan x with text prompts $\mathcal{V}$.

Details of network training

When pre-processing the data, all the chest CT images and CTVs were initially re-sampled to have an identical voxel spacing of $1.0 \times 1.0 \times 3.0$ mm^3. The image intensity values were truncated between −1000 and 1000 of Hounsfield unit, and linearly normalized within a range between 0 and 1.0. When training the network, a 3D patch with a size of $384 \times 384 \times 128$ pixels was randomly cropped to cover the entire breast alongside with its paired clinical data with batch size of 2. When evaluating the trained network, the entire 3D CT image was tested using sliding windows with a 3D patch with a size of $384 \times 384 \times 128$ pixels. We set the optimal hyper-parameters as listed in Supplementary Table 8. During training, we let the entire LLM frozen, while making the image encoder/decoder modules, the interactive alignment modules, and their corresponding linear layers, and the text prompts trainable parameters.

As the loss function, we computed both the binary CE loss and the Dice loss, with the weight value for each loss as 1.0, respectively. The network parameters were optimized using AdamW[40] optimizer with a learning rate of 0.0001 until the training epoch reaching 100. We implemented the network using the open-source library MONAI. All the experiments were conducted using the PyTorch[41] in Python using CUDA 11.4 on NVIDIA RTX A6000 48 GB. We further described backbones for each model, and compared training complexity in Supplementary Table 9.

Rationales of selecting baseline models

Our baselines, ConTEXTualNet[24], LISA[23], and HIPIE[22] along with our proposed model, LLMSeg, are designed to extract characteristics from an input sentence that are not explicitly visible in the image, as categorized in Supplementary Table 10. For example, tasks may include identifying the food item richest in Vitamin C from an image and generating a segmentation mask, or recognizing medical conditions and treatment plans (like cT2, N1mi, breast conserving surgery, and left-side procedures). These tasks necessitate a deep understanding of the sentence context and the ability to infer answers for context-aware

or reasoning/referring-based segmentation. Both ConTEXTualNet, LISA, and HIPIE, like our model, leverage text embeddings derived from a language model to facilitate multimodal segmentation.

Additionally, for a meaningful comparative study, it is crucial to retrain the baseline models with our 3D CT training data. ConTEXTualNet, being a CNN-based network designed for end-to-end training, allows us to adapt the original 2D model into a 3D model suitable for retraining with our 3D data. On the other hand, recent SOTA multimodal foundation models for segmentation, such as LISA[42] and HIPIE[22], utilize 2D SAM[16] or CLIP[27]-based cross-attention modules. Adapting these models to process 3D volumes as a whole would require retraining the 2D foundation model with 3D data, which is not feasible given our constraints. Consequently, to preserve their transfer learning mechanism based on the frozen 2D foundation model, we retrain these models by converting 3D CT scans to 2D slices as inputs. This highlights a limitation of current 2D vision-language models when adapting to 3D images, resulting in the loss of volumetric context for clinical information-guided multimodal segmentation and yielding suboptimal performance.

The reason for not including traditional open-vocabulary segmentation models in our study is that they are designed for semantic segmentation of visually discernible objects in an image, such as walls, chairs, windows, floors, and ceilings, as depicted in Supplementary Table 10. This capability stems from their use of pre-trained 2D vision-language foundation models which serves as their frozen backbone for feature extraction. These models leverage pre-aligned word-image features for semantic segmentation, thus, there are not appropriate baselines for our medical context-aware segmentation purposes, as the radiotherapy target volumes in CT images are not visually identifiable.

Details of evaluation

To quantitatively evaluate the CTV delineation performance, we calculated Dice coefficient (Dice), Intersection over Union (IoU), and the 95th percentile of Hausdorff Distance (95-HD)[43] to measure spatial distances between the ground-truth and the predicted contours. When calculating the 95-HD, all the measured distances in the pixel unit are converted with respect to the original pixel resolution, and the results are expressed in centimeters (cm).

Details of clinical evaluation

To accurately assess the performance of the model, we conducted clinical evaluations by the board-certified radiation oncologist with over 5 years of experience. To provide a more detailed evaluation of the model's performance and establish an objective criterion for assessment, we employed rubrics proposed by the radiation oncologists. For breast cancer, these rubrics included laterality (right, left, or bilateral−1 point), type of surgery (whether the case was post-BCS or mastectomy−1 point), volume definition (accurate definition of breast or chest wall, inclusion of regional LNs−1.5 points), coverage (ensuring the target volume was adequately covered without encompassing unnecessary areas), and integrity (absence of incomplete or distorted segmentation output), constituting a total of 5 points. Detailed criteria for each rubric and illustrative examples are provided in Supplementary Fig. 1 and Supplementary Table 1.

For prostate cancer, the criteria included primary site (accuracy in defining the treatment scope for the prostate, including seminal vesicles), volume definition (appropriate inclusion of the prostate and regional nodes), coverage, and integrity, totaling 4 points. The rubrics of laterality, surgery type, volume definition, and primary site were established to assess the appropriateness of the underlying concepts in defining the scope of the target area. Conversely, the criteria for coverage and integrity were specifically designed to evaluate the quality of the contouring. Detailed criteria for each rubric and illustrative examples are provided in Supplementary Fig. 2 and Supplementary Table 4.

Utilizing these evaluation criteria, to ensure fairness, the same board-certified radiation oncologists conducted assessments of the segmentation outputs by comparing them to the ground truth and considering the clinical context, all while being blinded to whether the outputs were generated by a vision-only model or a multimodal model.

Statistics and reproducibility

For statistical analysis, we used the non-parametric bootstrap method to calculate the confidence interval (CI) for each metric. We randomly sampled the total size of dataset from the original dataset while allowing replacement for 1000 times, repeatedly. Then, the mean values and the 95th percentile of CIs were estimated from the relative frequency distribution of each trial. Two-tailed Student's paired *t*-test was used for the statistical comparison between the two groups. No statistical method was used to predetermine sample size. No data were excluded from the analyses; The experiments were not randomized; The Investigator was not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Individual de-identified participant data cannot be freely shared due to privacy laws, specifically the Personal Information Protection Act of the Republic of Korea. Data sharing may be considered upon contacting the corresponding author, J.C.Y. (jong.ye@kaist.ac.kr), who will assess compliance with these legal provisions. If deemed appropriate, data sharing will proceed following formal inter-institutional collaboration agreements. Initial requests will receive a response within 10 working days. The source data to be shared includes specific outcome values of all patients used to generate the graphs and tables in this study. Instead of the complete patient dataset, a small sample of subjects with similar characteristics has been made available as open source for validation purposes at https://github.com/tvseg/MM-LLM-RO[44]. No additional documents, such as study protocols or statistical analysis plans, will be provided. While individual patient data will not be directly shared, the open-source sample data will remain available indefinitely. Data usage is restricted to research purposes only, and redistribution is prohibited. Source data is provided with this work. All remaining data is available in the manuscript, source data file, or supplementary information file. Source data are provided with this paper.

Code availability

The PyTorch codes for the proposed Multimodal AI used in this study is available at the following GitHub repository: https://github.com/tvseg/MM-LLM-RO[44].

References

1. Huynh, E. et al. Artificial intelligence in radiation oncology. *Nat. Rev. Clin. Oncol.* **17**, 771–781 (2020).
2. Shi, F. et al. Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. *Nat. Commun.* **13**, 6566 (2022).
3. Zhang, L. et al. Segment anything model (sam) for radiation oncology. arXiv preprint arXiv:2306.11730 (2023).
4. Chung, S. Y. et al. Clinical feasibility of deep learning-based auto-segmentation of target volumes and organs-at-risk in breast cancer patients after breast-conserving surgery. *Radiat. Oncol.* **16**, 1–10 (2021).

5. Offersen, B. V. et al. Estro consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. *Radiother. Oncol.* **114**, 3–10 (2015).
6. Choi, M. S. et al. Clinical evaluation of atlas-and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. *Radiother. Oncol.* **153**, 139–145 (2020).
7. Guo, Z., Guo, N., Gong, K. & Li, Q. et al. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. *Phys. Med. Biol.* **64**, 205015 (2019).
8. Liu, C. et al. Artificial general intelligence for radiation oncology. *Meta Radiol.* **1**, 100045 (2023).
9. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv preprint arXiv:2303.12712 (2023).
10. Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023).
11. Liu, Z. et al. Radiology-GPT: A large language model for radiology. arXiv preprint arXiv:2306.08666 (2023).
12. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
13. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
14. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., et al. Towards generalist biomedical AI. *NEJM AI* **1**, AIoa2300138 (2024).
15. Lee, S., Kim, W. J., Chang, J. & Ye, J. C. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. 2305.11490 (2024).
16. 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 *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026 (IEEE, 2023).
17. Kim, K., Oh, Y. & Ye, J. C. ZegOT: Zero-shot segmentation through optimal transport of text prompts. arXiv preprint arXiv:2301.12171 (2023).
18. Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B. & Lim, S.-N. Visual Prompt Tuning. In *Proc. 17th European Conference on Computer Vision (ECCV)*, 709–727 (Springer, 2022).
19. Zhou, K., Yang, J., Loy, C. C. & Liu, Z. Conditional prompt learning for vision-language models. 2203.05557 (2023).
20. Zhu, L., Chen, T., Ji, D., Ye, J. & Liu, J. Llafs: when large language models meet few-shot segmentation. 2311.16926 (2024).
21. Wang, W. et al. Visionllm: large language model is also an open-ended decoder for vision-centric tasks. 2305.11175 (2023).
22. Wang, X. et al. Hierarchical open-vocabulary universal image segmentation. *Adv. Neural Inform. Process. Syst.* **36**, (2024).
23. Lai, X. et al. LISA: reasoning segmentation via large language model. 2308.00692 (2023).
24. Huemann, Z. et al. ConTEXTual net: a multimodal vision-language model for segmentation of pneumothorax. *J. Imaging Inform. Med.* **1**, 1–12 (2024).
25. Hatamizadeh, A. et al. UNETR: transformers for 3d medical image segmentation. In *Proc. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 1748–1758 (IEEE, 2022).
26. Xing, Z., Ye, T., Yang, Y., Liu, G. & Zhu, L. Segmamba: long-range sequential modeling mamba for 3d medical image segmentation. 2401.13560 (2024).
27. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. International conference on machine learning*, 8748–8763 (PMLR, 2021).
28. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. In *Proc. International Conference on Learning Representations (ICLR)* (ICLR, 2022).
29. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017).
30. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
31. Rajpurkar, P. et al. ChexNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017).
32. Choi, B. G. et al. Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. *Yonsei Med. J.* **60**, 191–199 (2019).
33. Yoo, T. K. et al. Osteoporosis risk prediction for bone mineral density assessment of postmenopausal women using machine learning. *Yonsei Med. J.* **54**, 1321–1330 (2013).
34. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).
35. Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**, 1399–1406 (2022).
36. Moon, J. H., Lee, H., Shin, W., Kim, Y.-H. & Choi, E. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE J. Biomed. Health Inform.* **26**, 6070–6080 (2022).
37. Huang, Z., Zhang, X. & Zhang, S. Kiut: knowledge-injected u-transformer for radiology report generation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19809–19818 (IEEE, 2023).
38. Hosny, A. et al. Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. *Lancet Digit. Health* **4**, e657–e666 (2022).
39. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Proc. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, Proceedings, Part II 19*, 424–432 (Springer, 2016).
40. Kingma, D. & Ba, J. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)* (ICLR, 2015).
41. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inform. Process. Syst.* **32**, (2019).
42. Xu, J. et al. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2955–2966 (IEEE, 2023).
43. Crum, W. R., Camara, O. & Hill, D. L. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans. Med. Imaging* **25**, 1451–1461 (2006).
44. Oh, Y. et al. Llm-driven multimodal target volume contouring in radiation oncology. https://doi.org/10.5281/zenodo.12792278 (2024).

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant No. RS-2023-00242164 to S.P., and also supported by the NRF grant funded by the Korea government (MSIT) (No. RS-2024-00336454), (No. RS-2023-00262527) to J.C.Y., (No. 2022R1A2C2008623) to J.S.K., and (No. RS-2024-00345854) to Y.O. Additionally, this work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation) to J.C.Y., and by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare (No. HI20C1234) to J.S.K.

Author contributions

Y.O. designed the study, extended the code, conducted all experiments, analyzed data, and contributed to manuscript preparation. S.P. conceptualized the study, gathered and labeled the data, analyzed data, and also contributed to manuscript preparation. H.K.B., Y.C., and I.J.L. were responsible for data collection and manuscript preparation. J.S.K. and J.C.Y. provided supervision throughout the project, from conception to discussion, and assisted in preparing the manuscript.

Competing interests

J.S.K. is a shareholder and employee of Oncosoft Inc, which may benefit from the research results presented in this paper. This potential conflict of interest has been disclosed and managed according to institutional policies.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-53387-y.

Correspondence and requests for materials should be addressed to Jin Sung Kim or Jong Chul Ye.

Peer review information *Nature Communications* thanks Yejin Kim, Yaozong Gao, Danielle Bitterman, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.