Guideline-Informed MLLM Reasoning for Pathology-Aware Postoperative Prostate CTV Segmentation

Yinhao Wu [1] [iD]   YXW2120@MAVS.UTA.EDU
Hengrui Zhao[2] [iD]   HENGRUI.ZHAO@UTSOUTHWESTERN.EDU
Haiqing Li[1] [iD]   HXL9110@MAVS.UTA.EDU
Wenliang Zhong[1]   WXZ9204@MAVS.UTA.EDU
Hehuan Ma[1]   HEHUAN.MA@MAVS.UTA.EDU
Yuzhi Guo[1]   YUZHI.GUO@MAVS.UTA.EDU
Dan Nguyen[2]   DAN.NGUYEN@UTSOUTHWESTERN.EDU
Daniel Yang[2]   DANIEL.YANG@UTSOUTHWESTERN.EDU
Steve Jiang[2] [iD]   STEVE.JIANG@UTSOUTHWESTERN.EDU
Junzhou Huang[1] [iD]   JZHUANG@EXCHANGE.UTA.EDU
[1] *The University of Texas at Arlington, Arlington, TX, USA.*
[2] *UT Southwestern Medical Center, Dallas, TX, USA.*

Editors: Under Review for MIDL 2026

Abstract

Accurate segmentation of the Clinical Target Volume (CTV) is a critical prerequisite for precise radiotherapy planning, pursuing complete irradiation of microscopic disease while minimizing toxicity to surrounding healthy organs. However, achieving automated CTV segmentation remains highly challenging due to the invisible microscopic disease on planning CT and the necessity of incorporating clinical context into delineation decisions. Unlike previous methods that rely solely on visual features or coarse global text reasoning, we propose **ReaCT**, a unified framework that reformulates CTV segmentation as a multimodal reasoning task by explicitly integrating pathological information with visual context. Specifically, we introduce a Guideline-Informed LLM Agent designed to emulate the information-retrieval workflow of radiation oncologists. By distilling knowledge from authoritative clinical guidelines, this agent filters and structures lengthy pathology reports into a concise set of determinative pathological attributes, effectively bridging the semantic gap between unstructured clinical records and segmentation networks. Furthermore, we develop an Attribute-Specific MLLM Reasoner built upon a 3D residual U-Net that performs fine-grained spatial reasoning. By leveraging a sequence of attribute-specific query tokens, the model disentangles the distinct target implications of individual pathological attributes, enabling fine-grained anatomical alignment via multi-scale fusion using Two-Way Transformers. Experiments on a postoperative prostate cancer dataset demonstrate that ReaCT achieves state-of-the-art segmentation performance and exhibits strong robustness, with pronounced improvements under limited-annotation settings.

Keywords: Clinical Target Volume, Radiotherapy Planning, 3D Image Segmentation.

1. Introduction

Radiotherapy is one of the most common treatments for cancer, delivering radiation doses to the target volume while sparing surrounding healthy tissues (Bi et al., 2019; Liu et al., 2021). Achieving the optimal therapeutic effect relies on precise treatment planning, including the

delineation of the Clinical Target Volume (CTV) for microscopic tumor extensions (Lee et al., 2018; Balagopal et al., 2021). Different from Gross Tumor Volume (GTV) which usually have a distinct contrast on CT images, CTV can be invisible on planning CT images and its segmentation presents a formidable challenge. This difficulty is exacerbated in postoperative radical prostatectomy, where the surgical removal of the prostate and nearby tissues leaves a void in the target area, and the CTV boundaries are usually invisible. This necessitates complex reasoning from clinical context in addition to visual perception.

To facilitate consistent and reliable delineation, radiation oncologists typically need to integrate patient-specific pathological attributes derived from pathology reports alongside planning CT images in their decision-making process. These attributes are essential because they must be interpreted together with consensus clinical guidelines to determine the appropriate boundaries of the CTV (Jansen et al., 2000; Chang et al., 2007). For instance, different pathological stages determine whether only the proximal base of the seminal vesicles should be included, whereas confirmed seminal vesicle invasion mandates the inclusion of the entire seminal vesicle bed (Dal Pra et al., 2023). Consequently, distinct from conventional segmentation tasks, CTV segmentation inherently requires the incorporation of multimodal knowledge to reason about regions susceptible to microscopic metastases that are indistinguishable based on visual features alone.

Recently, Multimodal Large Language Models (MLLMs) have demonstrated exceptional reasoning capabilities when handling implicit or abstract text instructions (Lai et al., 2024; Zou et al., 2025). Drawing inspiration from these advancements and considering the intrinsic reasoning demands of CTV segmentation, we posit that empowering MLLMs to perform multimodal reasoning is essential for resolving the ambiguity of invisible target boundaries. However, seamlessly integrating such specialized knowledge into MLLM architectures presents significant challenges. First, patient records such as pathology reports and clinical notes are typically lengthy, unstructured, and rich in domain-specific terminology, making it difficult to effectively encode and align them with segmentation networks. Second, existing MLLM-based approaches typically rely on a single, coarse reasoning token, which limits their ability to capture the fine-grained correspondence between individual pathological attributes and their distinct spatial implications for CTV segmentation. A more comprehensive review of related work is provided in Appendix A.

To address the above limitations, we propose **ReaCT**, a unified multimodal framework that reformulates CTV segmentation as a reasoning task by integrating patient-specific pathological attributes with visual context. As illustrated in Figure 1, ReaCT employs an LLM agent to emulate the information-retrieval workflow of radiation oncologists. This agent first distills relevant clinical radiotherapy guidelines to derive a principled set of pathological attributes that determine the spatial extent of the CTV. Leveraging this distilled knowledge, the agent then processes raw pathology reports through a multi-stage pipeline involving keyword retrieval, semantic verification, and value standardization, transforming unstructured documentation into a concise and structured attribute set for downstream multimodal reasoning. Subsequently, ReaCT introduces a multimodal CTV segmentation network built upon a 3D residual U-Net backbone and a custom MLLM reasoner. The MLLM Reasoner jointly processes visual and textual tokens together with a sequence of attribute-specific <SEG> query tokens. This granular design enables fine-grained reasoning, where each query token independently encapsulates the distinct target implications of a

specific pathological attribute by aggregating relevant multimodal context. The hidden embeddings from the last layer corresponding to these `<SEG>` tokens are aggregated and then fused with multi-scale visual features through bi-directional transformer modules at each decoder stage to ensure precise anatomical alignment. Benefiting from this fine-grained multimodal fusion, ReaCT achieves state-of-the-art performance on postoperative prostate cancer datasets and exhibits pronounced robustness even in limited-annotation regimes. We highlight the following contributions:

- We propose **ReaCT**, a unified framework that reformulates CTV segmentation as a multimodal reasoning task by explicitly integrating patient-specific pathological attributes with visual context. This formulation addresses the inherent clinical need for multimodal integration and provides a principled mechanism toward anatomically and clinically coherent CTV segmentation.

- We design a Guideline-Informed LLM Agent that emulates the information-retrieval workflow of radiation oncologists. By distilling knowledge from consensus guidelines, the agent filters and structures lengthy pathology reports into a concise set of determinative attributes, effectively bridging the semantic gap between unstructured clinical records and segmentation networks.

- We develop an Attribute-Specific MLLM Reasoner that performs fine-grained spatial reasoning through a sequence of query tokens, enabling the model to disentangle the distinct target implications of individual pathological attributes and ensuring precise anatomical alignment even in limited-annotation regimes.

2. Methodology

As illustrated in Figure 1, ReaCT comprises two parts. First, a Guideline-Informed LLM Agent emulates the expert information-retrieval workflow, distilling consensus guidelines to transform unstructured pathology reports into a concise set of determinative attributes. Second, a multimodal CTV segmentation network predicts the CTV mask by integrating the 3D CT volume with patient-specific attributes. Built upon a 3D U-Net and an Attribute-Specific MLLM Reasoner, ReaCT utilizes a sequence of query tokens to generate fine-grained spatial reasoning embeddings, which are fused into the decoder via multi-scale Two-Way transformers to ensure precise anatomical alignment.

2.1. Guideline-Informed LLM Agent

To bridge the semantic gap between unstructured clinical records and the downstream segmentation network, we construct a Guideline-Informed LLM Agent based on GPT-4o (Hurst et al., 2024). This agent emulates the information-retrieval workflow of radiation oncologists through a two-step process: (1) distilling relevant knowledge from consensus guidelines for radiotherapy into a compact attribute schema, and (2) extracting patient-specific attribute values from raw pathology reports. Detailed implementation is provided in Appendix B.

2.1.1. Guideline-Based Schema Construction

To establish a grounded reasoning logic, the agent first constructs a comprehensive guideline corpus $\mathcal{G} = \{g_m\}_{m=1}^M$ by retrieving and aggregating consensus guidelines for radiotherapy,
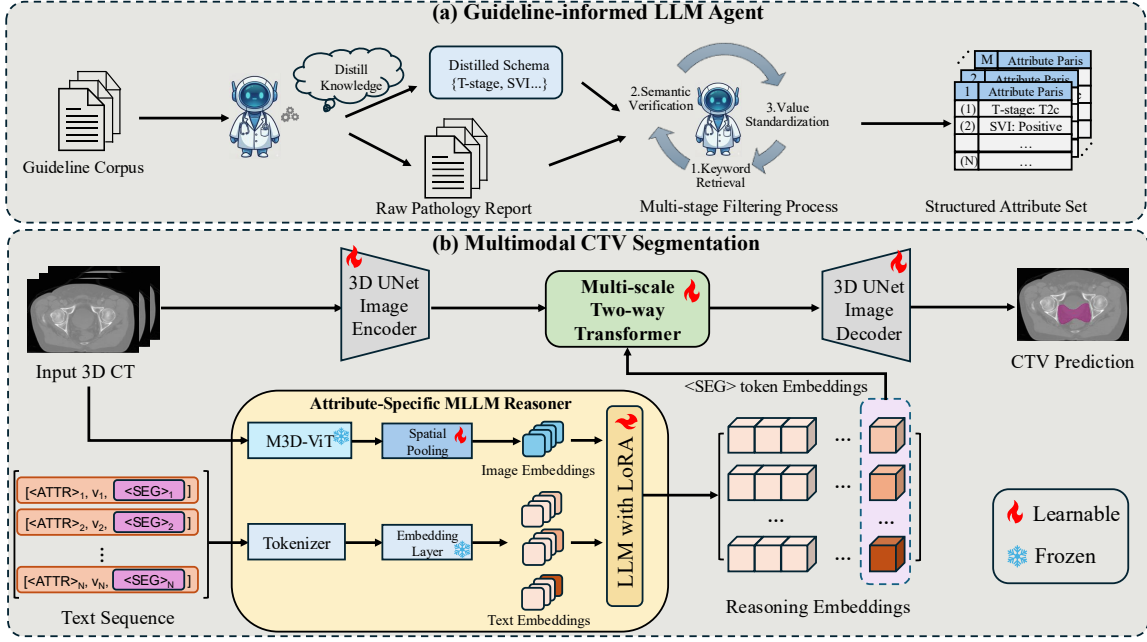
Figure 1: The overall framework of ReaCT. (a) An LLM Agent emulates expert workflow by distilling knowledge from consensus guidelines to obtain structured determinative attributes. (b) A Multimodal CTV Segmentation Network integrates a 3D U-Net with an Attribute-Specific MLLM Reasoner, fusing fine-grained reasoning embeddings derived from distinct query tokens into the decoder via multi-scale Two-Way Transformers to ensure precise anatomical alignment.

such as ESTRO (Niyazi et al., 2016), NCCN (Carroll et al., 2016), and RTOG(Kruser et al., 2019). Subsequently, under a fixed prompting policy $\mathcal{P}_{\text{schema}}(\cdot)$ with deterministic decoding, the agent employs the LLM Ψ_{LLM} to analyze these retrieved documents and distill the pathological factors that explicitly influence the CTV definition. Formally, this retrieval and distillation process yields a principled attribute schema $\mathcal{S} = \Psi_{\text{LLM}}\big(\mathcal{P}_{\text{schema}}(\mathcal{G})\big) = \{k_i\}_{i=1}^{N}$, where each k_i corresponds to a clinically determinative pathological attribute.

2.1.2. Multi-Stage Attribute Extraction.

Guided by the schema $\mathcal{S}$, the agent processes the raw patient-specific pathology report $\mathbf{D}_{\text{raw}}$ to derive a structured set of attribute-value pairs. To mitigate the noise and redundancy inherent in unstructured medical text, the agent performs a coarse-to-fine filtering pipeline. First, it applies a keyword matching operator $\mathcal{F}_{\text{key}}(\cdot)$ to retrieve a subset of relevant text segments $\mathbf{T}_{\text{rel}} \subset \mathbf{D}_{\text{raw}}$ potentially containing information linked to $\mathcal{S}$, thereby efficiently narrowing the search space. Subsequently, the agent functions as a semantic verifier $\Phi_{\text{verify}}(\cdot)$ on $\mathbf{T}_{\text{rel}}$. It filters out irrelevant narratives (e.g., unrelated medical history) and eliminates redundant statements, retaining only textual evidence that directly informs the attributes

4

in $\mathcal{S}$. Finally, the agent aggregates the verified context to assign a standardized value v_k for each attribute k. This produces a structured clinical attribute set $\mathcal{A} = \{(k, v_k) \mid k \in \mathcal{S}, v_k \neq \varnothing\}$. The curated attribute set $\mathcal{A}$ serves as textual input for the multimodal reasoning, ensuring more accurate and context-aware CTV segmentation.

2.2. Multimodal CTV Segmentation

The segmentation network consists of a 3D U-Net-based segmentation path that encodes and decodes multi-scale spatial features, and an MLLM reasoner that generates fine-grained reasoning embeddings from the clinical attributes and corresponding image context.

2.2.1. SEGMENTATION PATH

We adopt a 3D residual U-Net (Çiçek et al., 2016) as the visual backbone to facilitate the hierarchical interaction between anatomical features and clinical reasoning. The encoder $\mathcal{F}_\phi$ first processes the input CT volume $\mathbf{x}_{\text{img}} \in \mathbb{R}^{B \times 1 \times H \times W \times D}$ to extract a pyramid of multi-scale spatial features $\{\mathbf{f}_l\}_{l=1}^L$, where $\mathbf{f}_l \in \mathbb{R}^{B \times C_l \times H_l \times W_l \times D_l}$ denotes the feature map at the l-th resolution scale. Meanwhile, the proposed MLLM Reasoner jointly encodes the visual tokens and the extracted pathological attributes to generate a set of fine-grained reasoning embeddings $\mathbf{H}_{\text{reason}} = \{\mathbf{h}_k\}_{k=1}^N \in \mathbb{R}^{B \times N \times d'}$, where each $\mathbf{h}_k$ captures the inferred spatial implication of the k-th pathological attribute. To integrate the fine-grained reasoning embeddings $\mathbf{H}_{\text{reason}}$ generated by the MLLM into the decoding path, we employ the Two-Way Transformer fusion mechanism from the Segment Anything Model (SAM) (Kirillov et al., 2023). This module facilitates bidirectional interaction between the spatial features $\mathbf{f}_l$ and attribute queries $\mathbf{H}_{\text{reason}}$ at each upsampling stage, yielding a clinically modulated representation $\tilde{\mathbf{f}}_l = \text{TwoWayBlock}(\mathbf{f}_l, \mathbf{H}_{\text{reason}})$. Finally, the decoder $\mathcal{D}_\phi$ progressively upsamples and aggregates these fused features $\{\tilde{\mathbf{f}}_l\}_{l=1}^L$ to reconstruct the segmentation mask, outputting the final CTV probability map $\hat{\mathbf{y}} \in [0, 1]^{B \times 1 \times H \times W \times D}$.

2.2.2. ATTRIBUTE-SPECIFIC MLLM REASONER

We design the MLLM reasoner to jointly encode 3D image features and the extracted pathological attributes, enabling reasoning-guided CTV segmentation. Specifically, we build upon M3D-LaMed (Bai et al., 2024), a specialized multimodal large language model for 3D medical imaging, which consists of a 3D ViT image encoder (M3D-ViT), a spatial pooling projector, and a LLaMA-2-7B (Touvron et al., 2023) language backbone.

Multimodal LLM for Fine-Grained Anatomical Reasoning. Given the 3D CT volume $\mathbf{x}_{\text{img}}$, the pretrained M3D-ViT encoder Φ_{ViT} first extracts patch-level embeddings $\mathbf{Z}_{\text{img}} = \Phi_{\text{ViT}}(\mathbf{x}_{\text{img}}) \in \mathbb{R}^{M_0 \times d_v}$, where M_0 denotes the number of 3D patches and d_v represents the vision hidden dimension. To align with the LLM latent space, these embeddings are processed by a spatial pooling projector $\mathcal{P}_\psi(\cdot)$, which applies 3D average pooling followed by a series of Multi-Layer Perceptrons (MLPs). This projection yields the compressed visual embedding $\mathbf{F}_{\text{img}} = \mathcal{P}_\psi(\mathbf{Z}_{\text{img}}) \in \mathbb{R}^{M \times d'}$, where $M = 128$ is the reduced token length and d' aligns with the hidden dimension of the LLM.

To achieve fine-grained anatomical reasoning, we augment the tokenizer's vocabulary with a set of learnable attribute-specific tokens $\{\texttt{<SEG>}_i\}_{i=1}^N$. Each token $\texttt{<SEG>}_i$ is designed

to act as a dedicated reasoning query for the i-th attribute in $\mathcal{A}$. Formally, let $\mathbf{A}_i$ denote the tokenized embedding of the attribute pair (k_i, v_i) (e.g., [SVI, Positive]). We construct the joint multimodal input sequence $\mathbf{X}$ by concatenating the visual context with a series of attribute-reasoning blocks. Formally, for the i-th attribute, the input sequence $\mathbf{X}_i$ is:

$$\mathbf{X}_i = \big[\mathbf{F}_{\text{img}}; \mathbf{A}_i; \texttt{<SEG>}_i\big]. \tag{1}$$

Each input sequence functions as an independent reasoning unit, prompting the model to synthesize the shared visual context $\mathbf{F}_{\text{img}}$ with the specific pathological attribute $\mathbf{A}_i$ to encode the spatial intent into the corresponding $\texttt{<SEG>}_i$ token.

Subsequently, the input sequence $\mathbf{X}$ is processed through the LLM backbone $\mathcal{M}_\theta$ with L transformer layers to model the conditional dependency $p(x_t|x_{<t})$, where t denotes the token index. The hidden states are recursively transformed as:

$$\mathcal{H}^{(\ell)} \; = \; \mathcal{M}_\theta^{(\ell)}\Big(\mathcal{H}^{(\ell-1)}\Big) = \big\{\, \mathbf{h}_1^{(\ell)}, \ldots, \mathbf{h}_T^{(\ell)} \,\big\}, \quad \ell = 1, \ldots, L. \tag{2}$$

To distill the fine-grained reasoning result, let t_i^* denote the position index of the i-th attribute-specific query token $\texttt{<SEG>}_i$ within the sequence. We extract the last-layer hidden state at this specific position to obtain the final reasoned embedding $\mathbf{h}_{t_i^*}^{(L)}$. Aggregating these embeddings over all N attributes yields the reasoning set $\mathbf{H}_{\text{reason}} = \{\mathbf{h}_{t_i^*}^{(L)}\}_{i=1}^N$, which provides disentangled, spatially-aware guidance for the segmentation decoder.

LoRA-based Adaptation. To efficiently adapt the pretrained $\mathcal{M}_\theta$ to CTV segmentation, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) on the query and value projections. Instead of full-parameter updates, the weight transformation is parameterized as $W' = W + BA$, where $A \in \mathbb{R}^{r \times d'}$ and $B \in \mathbb{R}^{d' \times r}$ represent learnable low-rank matrices ($r \ll d'$). This strategy minimizes computational overhead while enabling the model to effectively capture the attribute correspondences essential for CTV segmentation.

2.3. Training Objective

We adopt a weighted combination of Dice loss and binary cross-entropy (BCE) to optimize the network, formulated as:

$$\mathcal{L}_{\text{total}} = \lambda_0 \, \mathcal{L}_{\text{Dice}}(\hat{\mathbf{y}}_i, \mathbf{y}_i) + \lambda_1 \, \mathcal{L}_{\text{BCE}}(\hat{\mathbf{y}}_i, \mathbf{y}_i), \tag{3}$$

where $\hat{\mathbf{y}}_i$ and $\mathbf{y}_i$ denote the predicted and ground-truth CTV masks, respectively, and λ_0, λ_1 are weighting coefficients. The Dice and BCE loss are formulated as $\mathcal{L}_{\text{Dice}} = 1 - \dfrac{2\sum_i \hat{\mathbf{y}}_i \mathbf{y}_i + \epsilon}{\sum_i \hat{\mathbf{y}}_i + \sum_i \mathbf{y}_i + \epsilon}$ and $\mathcal{L}_{\text{BCE}} = -\dfrac{1}{|\Omega|} \sum_{j \in \Omega} \left[\mathbf{y}_j \log \hat{\mathbf{y}}_j + (1 - \mathbf{y}_j) \log(1 - \hat{\mathbf{y}}_j)\right]$, where $|\Omega|$ is the number of voxels and ϵ is a small constant for numerical stability.

3. Experiments

3.1. Datasets and Implementation Details

We conduct experiments on a large-scale multimodal in-house dataset comprising 688 post-operative prostate cancer patients, collected from the Department of Radiation Oncology

at UT Southwestern Medical Center. This cohort represents a clinically demanding scenario where the prostate has been surgically removed, requiring the CTV to be inferred by synthesizing surrounding anatomical landmarks (e.g., bladder, rectal wall) with patient-specific pathological attributes. Ground-truth CTV masks were manually delineated by six experienced radiation oncologists following consensus guidelines. In addition, the operative pathology report associated with each patient includes critical pathological attributes, such as pathological T-stage, Gleason Score, and Seminal Vesicle Invasion (SVI) status. To ensure rigorous evaluation, we perform five randomized training-validation splits. Specifically, in each iteration, approximately 90% ($N_{\text{train}} = 496$) of the data is allocated for training and 10% ($N_{\text{val}} = 54$) for validation. Furthermore, 138 cases are reserved as a fixed hold-out test set to assess the final performance. All data splitting is performed at the patient level to prevent data leakage.

All CT volumes are preprocessed following the nnU-Net (Isensee et al., 2021) pipeline, including isotropic resampling to $1 \times 1 \times 1$ mm^3, intensity normalization, and foreground cropping. The model is implemented in MONAI (Cardoso et al., 2022) with a patch size of $320 \times 320 \times 64$ and batch size 1. The M3D-ViT input resolution is $32 \times 256 \times 256$, consistent with its pre-training setup, and the maximum LLM context length is 512 with 128 visual tokens. LoRA fine-tuning is applied to the query and value projections (q_proj, v_proj) using rank $r = 16$, scaling factor $\alpha = 16$, and dropout 0.05. Standard data augmentation (e.g., rotation, scaling, flipping) is used during training. The network is optimized with AdamW (Loshchilov and Hutter, 2019) (lr=1×10^{-5}, weight decay=1×10^{-8}) for up to 50 epochs on a cluster equipped with six NVIDIA H100 GPUs. Performance is evaluated using Dice Similarity Coefficient (Dice), 95th Percentile Hausdorff Distance (HD95), and Average Symmetric Surface Distance (ASSD).

3.2. Comparative Results

To evaluate the segmentation accuracy of ReaCT, we compare it with both vision-only and multimodal segmentation baselines. As summarized in Table 1, ReaCT achieves state-of-the-art performance, with a Dice score of 0.8185, HD95 of 4.15 mm, and ASSD of 1.38 mm. First, vision-only baselines (e.g., 3D U-Net, nnU-Net) achieve relatively stable Dice results; however, the HD95 and ASSD metrics still indicate a deficiency in identifying precise CTV boundaries. This highlights the importance of pathological information to provide critical guidance for regions that are radiographically indistinguishable based on imaging modalities alone. Second, text-prompted segmentation methods, including BiomedParse, SAT, and Medformer, show worse performance than vision-only baselines. We attribute this to their reliance on text encoders with encoder-only architectures. While these text encoders excel at processing explicit, content-descriptive prompts, they lack the reasoning capacity required to translate abstract and implicit pathological attributes into effective segmentation embeddings. Consequently, while their fixed prompt templates allow for flexible text-conditioned segmentation, they fall short in the deep clinical reasoning essential for CTV segmentation. Furthermore, compared to LLMSeg, which similarly employs LLaMA-2 to combine electronic medical records with images, ReaCT incorporates a more comprehensive set of attributes to enable fine-grained reasoning in postoperative scenarios where the GTV has been surgically removed. Moreover, ReaCT adopts an MLLM reasoner that jointly pro-

Table 1: Quantitative results of CTV segmentation on the in-house dataset (Mean ± SD). ↑: higher is better; ↓: lower is better.

Methods	Dice ↑	HD95 (mm) ↓	ASSD (mm) ↓
3D U-Net (Çiçek et al., 2016)	$0.7847_{\pm 0.01}$	$6.97_{\pm 2.33}$	$2.13_{\pm 0.54}$
nnU-Net (Isensee et al., 2021)	$0.7822_{\pm 0.01}$	$11.69_{\pm 4.31}$	$3.70_{\pm 0.85}$
UNETR (Hatamizadeh et al., 2022)	$0.7843_{\pm 0.01}$	$7.02_{\pm 2.20}$	$2.15_{\pm 0.55}$
Swin-UNETR (Hatamizadeh et al., 2021)	$0.7965_{\pm 0.01}$	$5.51_{\pm 1.25}$	$1.88_{\pm 0.44}$
U-Mamba (Ma et al., 2024)	$0.7715_{\pm 0.02}$	$7.50_{\pm 1.80}$	$2.25_{\pm 0.55}$
BiomedParse (Zhao et al., 2025a)	$0.7680_{\pm 0.04}$	$8.85_{\pm 1.50}$	$2.45_{\pm 0.45}$
SAT (Zhao et al., 2025b)	$0.7560_{\pm 0.03}$	$9.20_{\pm 2.20}$	$2.68_{\pm 0.70}$
Medformer (Rajendran et al., 2025)	$0.7750_{\pm 0.06}$	$9.80_{\pm 3.10}$	$2.80_{\pm 0.85}$
LLMSeg (Oh et al., 2024)	$0.7857_{\pm 0.01}$	$5.20_{\pm 1.02}$	$1.75_{\pm 0.44}$
w/o Textual Tokens	$0.8015_{\pm 0.05}$	$4.52_{\pm 0.75}$	$1.51_{\pm 0.18}$
w/o Visual Tokens	$0.7942_{\pm 0.04}$	$5.05_{\pm 0.85}$	$1.64_{\pm 0.20}$
w/o MLLM Reasoner	$0.7885_{\pm 0.02}$	$5.85_{\pm 1.10}$	$1.92_{\pm 0.40}$
Generic Text Prompt	$0.7989_{\pm 0.04}$	$4.98_{\pm 0.80}$	$1.63_{\pm 0.19}$
Concat Attributes	$0.8017_{\pm 0.05}$	$4.65_{\pm 0.70}$	$1.56_{\pm 0.15}$
ReaCT (Ours)	$\mathbf{0.8185}_{\pm 0.05}$	$\mathbf{4.15}_{\pm 1.66}$	$\mathbf{1.38}_{\pm 0.48}$

cesses visual and textual tokens, which promotes more effective reasoning than relying on textual information only. Qualitative visualizations provided in the Appendix B.3.1 further demonstrate that ReaCT produces contours with better anatomical consistency, especially in challenging regions where boundaries are ambiguous.

3.3. Ablation Results

3.3.1. IMPACT OF MULTIMODAL REASONING COMPONENTS.

We conduct ablation studies to validate the contributions of the MLLM-based reasoner and the specific prompt design strategies. First, to examine the role of each modality, we evaluate variants using only image tokens (*w/o Textual Tokens*) or only attribute text (*w/o Visual Tokens*) as input to the MLLM. As shown in Table 1, both variants exhibit clear performance drops compared to ReaCT, indicating that neither modality alone is sufficient for accurate CTV segmentation. Notably, the *w/o Textual Tokens* variant (Dice: 0.8015) still outperforms standard vision-only baselines, suggesting that the MLLM backbone enhances the expressiveness of visual representations even without explicit textual guidance. Conversely, the *w/o Visual Tokens* variant performs significantly worse (Dice: 0.7942), confirming that pathological attributes alone cannot resolve anatomical details without visual context. Furthermore, replacing the MLLM reasoner with a standard biomedical text encoder (PubMedBERT (Gu et al., 2021)) followed by Two-Way Transformer fusion (*w/o MLLM Reasoner*) leads to a substantial degradation (Dice: 0.7885). This demonstrates that encoder-only text encoders are insufficient for capturing the conditional reasoning required for this task, whereas the MLLM provides essential joint reasoning capabilities.
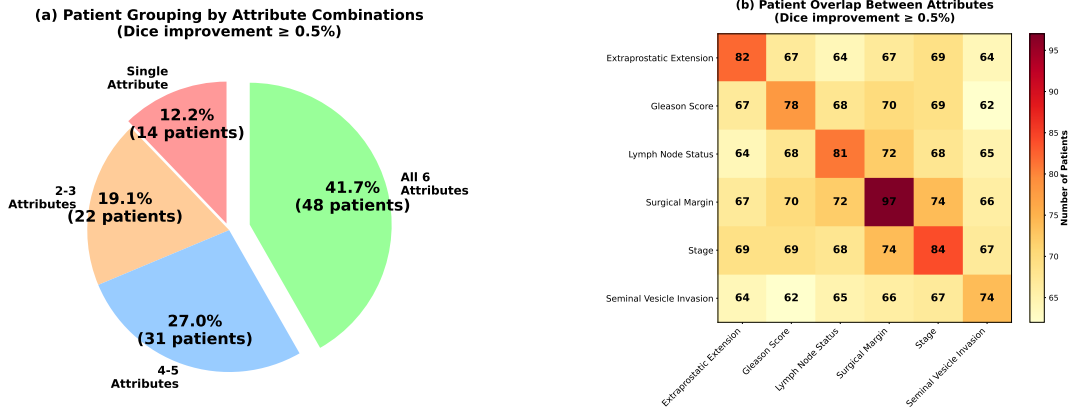
Figure 2: Analysis of Clinical Attribute Efficacy. (a) shows that most cases benefit from multi-attribute reasoning, while (b) reveals the biological correlations and complementary nature of these pathological attributes.

3.3.2. Effectiveness of Fine-Grained Attribute Reasoning

To verify the necessity of our fine-grained, multi-token design, we compare ReaCT against two alternative prompting strategies: (1) Generic Text Prompt, which uses a static instruction (i.e., *Segment the postoperative Clinical Target Volume for prostate cancer based on the CT image*); and (2) Concat Attributes, which concatenates all extracted attributes into a single sequence followed by a single unified `<SEG>` query token. The Generic Text Prompt yields a Dice score of 0.7989, which indicates that without patient-specific context, generic instructions fail to provide effective guidance for CTV segmentation. The Concat Attributes strategy improves performance to 0.8017, yet it still lags significantly behind ReaCT. This result supports our hypothesis that compressing diverse pathological factors into a single global representation creates a semantic bottleneck, preventing the model from disentangling their distinct spatial implications. In contrast, ReaCT's use of a sequence of attribute-specific query tokens enables the model to explicitly reason about how each attribute dictates local boundaries, leading to improved segmentation accuracy. Appendix B.3.2 further confirms this reliance on active semantic reasoning, where deliberately corrupting attribute values leads to significant performance degradation.

3.3.3. Impact of Individual and Combinatorial Attributes.

To further validate the necessity of each extracted attribute, we analyze the patient cohort where ReaCT yields significant segmentation improvements (i.e. ΔDice $\geq 0.5\%$). As shown in Figure 2(a), the majority of patients (87.8%) benefit from the integration of multiple clinical attributes rather than single ones, among the six pathological attributes used in this study. This highlights the complementary nature of clinical information and the heterogeneity of patient-specific treatment responses. Furthermore, the attribute overlap matrix in Figure 2(b) reveals strong inter-attribute correlations, suggesting that different pathological factors capture overlapping yet distinct aspects of tumor characteristics. For
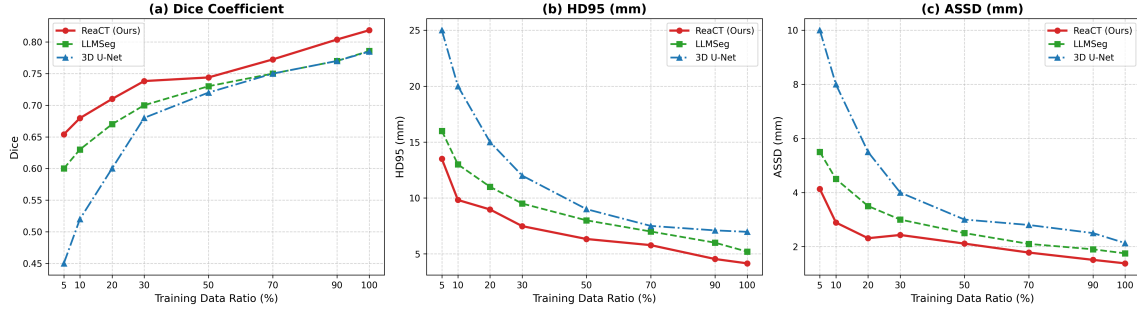
Figure 3: Performance comparison of ReaCT, 3D U-Net, and LLMSeg across varying training data ratios (5%–100%). ReaCT consistently demonstrates superior performance even in severe low-data regimes.

instance, high Gleason scores frequently co-occur with positive lymph node status, reflecting the known biological propensity for high-grade tumors to metastasize. Notably, Surgical Margin, Pathological Stage, and Extraprostatic Extension emerge as the most influential factors, showing the highest diagonal densities. This is consistent with clinical practice, where these parameters serve as key determinants in defining CTV expansion boundaries.

3.3.4. Data Efficiency and Robustness Analysis

To investigate sample efficiency, we evaluate the performance of ReaCT against a representative vision-only baseline (3D U-Net) and a multimodal baseline (LLMSeg) under varying training data proportions ranging from 5% to 100%. As illustrated in Figure 3, ReaCT consistently outperforms both baselines across all data regimes. Notably, the 3D U-Net suffers severe degradation in extreme low-data settings (e.g., 5%–20%), as indicated by a sharp drop in metrics. This confirms that without semantic guidance, visual features alone are insufficient to generalize from sparse supervision. While LLMSeg exhibits better stability than the vision-only model, it still lags behind ReaCT. These results highlight that explicitly modeling pathological reasoning substantially improves label efficiency and robustness, even when pixel-level supervision is scarce.

4. Conclusion

In this work, we present **ReaCT**, a unified framework reformulating CTV segmentation as a multimodal reasoning task. By utilizing a Guideline-Informed LLM Agent to distill determinative attributes and an Attribute-Specific MLLM Reasoner for fine-grained spatial inference, ReaCT effectively bridges the gap between abstract clinical logic and invisible anatomical boundaries. Experiments on a large-scale prostate dataset demonstrate state-of-the-art performance and remarkable robustness in limited-annotation regimes. Future work will focus on adaptively integrating evolving clinical guidelines to enhance generalization across diverse disease sites and institutional standards.

References

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Anjali Balagopal, Dan Nguyen, Howard Morgan, Yaochung Weng, Michael Dohopolski, Mu-Han Lin, Azar Sadeghnejad Barkousaraie, Yesenia Gonzalez, Aurelie Garant, Neil Desai, et al. A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. *Medical image analysis*, 72:102101, 2021.

Nan Bi, Jingbo Wang, Tao Zhang, Xinyuan Chen, Wenlong Xia, Junjie Miao, Kunpeng Xu, Linfang Wu, Quanrong Fan, Luhua Wang, et al. Deep learning improved clinical target volume contouring quality and efficiency for postoperative radiation therapy in non-small cell lung cancer. *Frontiers in oncology*, 9:1192, 2019.

Carlos E. Cardenas, Brian M. Anderson, Michalis Aristophanous, Jinzhong Yang, Dong Joo Rhee, Rachel E McCarroll, Abdallah Sherif Radwan Mohamed, Mona Kamal, Baher A. Elgohari, Hesham M. Elhalawani, Clifton David Fuller, Arvind U. K. Rao, Adam S. Garden, and Laurence E. Court. Auto-delineation of oropharyngeal clinical target volumes using 3d convolutional neural networks. *Physics in Medicine & Biology*, 63, 2018. URL https://api.semanticscholar.org/CorpusID:53226253.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Peter R Carroll, J Kellogg Parsons, Gerald Andriole, Robert R Bahnson, Erik P Castle, William J Catalona, Douglas M Dahl, John W Davis, Jonathan I Epstein, Ruth B Etzioni, et al. Nccn guidelines insights: prostate cancer early detection, version 2.2016. *Journal of the National Comprehensive Cancer Network*, 14(5):509–519, 2016.

Eric L Chang, Serap Akyurek, Tedde Avalos, Neal Rebueno, Chris Spicer, John Garcia, Robin Famiglietti, Pamela K Allen, KS Clifford Chao, Anita Mahajan, et al. Evaluation of peritumoral edema in the delineation of radiotherapy clinical target volumes for glioblastoma. *International Journal of Radiation Oncology* Biology* Physics*, 68(1): 144–150, 2007.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

Alan Dal Pra, Piet Dirix, Vincent Khoo, Christian Carrie, Cesare Cozzarini, Valérie Fonteyne, Pirus Ghadjar, Alfonso Gomez-Iturriaga, Valeria Panebianco, Almudena Zapatero, et al. Estro acrop guideline on prostate bed delineation for postoperative radiotherapy in prostate cancer. *Clinical and translational radiation oncology*, 41:100638, 2023.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Edwin PM Jansen, Luc GH Dewit, Marcel van Herk, and Harry Bartelink. Target volumes in radiotherapy for high-grade malignant glioma of the brain. *Radiotherapy and Oncology*, 56(2):151–156, 2000.

Dakai Jin, Dazhou Guo, Tsung-Ying Ho, Adam P Harrison, Jing Xiao, Chen-Kan Tseng, and Le Lu. Deeptarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. *Medical Image Analysis*, 68:101909, 2021.

Kwanyoung Kim, Yujin Oh, Sangjoon Park, Hwa Kyung Byun, Joongyo Lee, Jin Sung Kim, Yong Bae Kim, and Jong Chul Ye. End-to-end breast cancer radiotherapy planning via lmms with consistency embedding. *Medical Image Analysis*, page 103646, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

Tim J Kruser, Walter R Bosch, Shahed N Badiyan, Joseph A Bovi, Amol J Ghia, Michelle M Kim, Abhishek A Solanki, Sean Sachdev, Christina Tsien, Tony JC Wang, et al. Nrg brain tumor specialists consensus guidelines for glioblastoma contouring. *Journal of neuro-oncology*, 143:157–166, 2019.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

Anne W Lee, Wai Tong Ng, Jian Ji Pan, Sharon S Poh, Yong Chan Ahn, Hussain AlHussain, June Corry, Cai Grau, Vincent Grégoire, Kevin J Harrington, et al. International guideline for the delineation of the clinical target volumes (ctv) for nasopharyngeal carcinoma. *Radiotherapy and Oncology*, 126(1):25–36, 2018.

Zhikai Liu, Wanqi Chen, Hui Guan, Hongnan Zhen, Jing Shen, Xia Liu, An Liu, Richard Li, Jianhao Geng, Jing You, et al. An adversarial deep-learning-based model for cervical cancer ctv segmentation with multicenter blinded randomized controlled validation. *Frontiers in Oncology*, 11:702270, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

Maximilian Niyazi, Michael Brada, Anthony J Chalmers, Stephanie E Combs, Sara C Erridge, Alba Fiorentino, Anca L Grosu, Frank J Lagerwaard, Giuseppe Minniti, René-Olivier Mirimanoff, et al. Estro-acrop guideline "target delineation of glioblastomas". *Radiotherapy and oncology*, 118(1):35–42, 2016.

Yujin Oh, Sangjoon Park, Hwa Kyung Byun, Yeona Cho, Ik Jae Lee, Jin Sung Kim, and Jong Chul Ye. Llm-driven multimodal target volume contouring in radiation oncology. *Nature Communications*, 15(1):9186, 2024.

Praveenbalaji Rajendran, Yong Yang, Thomas R Niedermayr, Michael Gensheimer, Beth Beadle, Quynh-Thu Le, Lei Xing, and Xianjin Dai. Large language model-augmented auto-delineation of treatment target volume in radiation therapy. *Arxiv*, pages arXiv–2407, 2024.

Praveenbalaji Rajendran, Yizheng Chen, Liang Qiu, Thomas Niedermayr, Wu Liu, Mark Buyyounouski, Hilary Bagshaw, Bin Han, Yong Yang, Nataliya Kovalchuk, et al. Autodelineation of treatment target volume for radiation therapy using large language model-aided multimodal learning. *International Journal of Radiation Oncology* Biology* Physics*, 121(1):230–240, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Fan Wang, Xuanang Xu, Defu Yang, Ronald C Chen, Trevor J Royce, Andrew Wang, Jun Lian, and Chunfeng Lian. Dynamic cross-task representation adaptation for clinical targets co-segmentation in ct image-guided post-prostatectomy radiotherapy. *IEEE transactions on medical imaging*, 42(4):1046–1055, 2022.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, 22(1):166–176, 2025a.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Xiao Zhou, Ya Zhang, Yanfeng Wang, and Weidi Xie. Large-vocabulary segmentation for medical images with text prompts. *NPJ Digital Medicine*, 8(1):566, 2025b.

Ke Zou, Yang Bai, Bo Liu, Yidi Chen, Zhihao Chen, Yang Zhou, Xuedong Yuan, Meng Wang, Xiaojing Shen, Xiaochun Cao, et al. Uncertainty-aware medical diagnostic phrase identification and grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Appendix A. Related Work

A.1. CTV Segmentation Methods

Compared to conventional segmentation tasks driven by visual contrast, Clinical Target Volume (CTV) segmentation requires identifying microscopic spread that is invisible on standard imaging. To bridge this visibility gap, previous methods have attempted to infer target boundaries using auxiliary geometric cues or handcrafted anatomical heuristics. For example, Cardenas *et al.* (Cardenas et al., 2018) proposed a dual-channel 3D U-Net that ingests both CT scans and Gross Tumor Volume (GTV) masks to infer target boundaries based on spatial proximity. Similarly, Jin *et al.* (Jin et al., 2021) introduced a framework incorporating signed distance maps from the GTV and adjacent organs to provide explicit geometric constraints. Specific to postoperative prostate cancer, Wang *et al.* (Wang et al., 2022) modeled the prostate bed as a virtual target to guide segmentation in the absence of the primary organ. While these approaches improve consistency, their reliance on fixed geometric heuristics or spatial expansions limits their adaptability to anatomical variations, particularly in postoperative scenarios where the tumor has been surgically removed. In contrast, ReaCT addresses the intrinsic need for multimodal integration in CTV segmentation, enabling clinically informed reasoning about target extent that aligns with the decision-making process of radiation oncologists.

A.2. LLM-based CTV Segmentation Methods

The integration of Large Language Models (LLMs) into radiotherapy workflows marks a significant shift towards utilizing clinical data as auxiliary information for target volume segmentation. For instance, LLMSeg (Oh et al., 2024) demonstrated the capability of LLMs to enhance CTV segmentation by encoding clinical texts, such as tumor stage and surgery type, for breast and prostate cancer. Building on this, RO-LMM (Kim et al., 2025) proposed a comprehensive agent covering tasks from report summarization to plan-guided segmentation, while Medformer (Rajendran et al., 2024, 2025) leveraged hierarchical vision transformers fused with LLM-extracted text features to improve target delineation. However, these existing methods largely treat LLMs as static text encoders that offer only coarse global conditioning, without exploiting their reasoning capacity to model how individual pathological factors influence local anatomical boundaries. Consequently, they fail to deliver the fine-grained, attribute-specific reasoning required for accurate CTV segmentation. In contrast, ReaCT introduces a guideline-informed attribute schema and an attribute-specific multimodal LLM that performs fine-grained reasoning over visual and textual cues, enabling clinically coherent and anatomically precise boundary prediction.

Appendix B. Details of Guideline-Informed LLM Agent

In this section, we provide the detailed prompt designs and workflow specifications for the Guideline-Informed LLM Agent. The agent is built upon GPT-4o and is engaged via a multi-stage cognitive workflow to emulate the information-retrieval process of a radiation oncologist. This process encompasses guideline knowledge distillation, relevant context retrieval, semantic verification, and value standardization.

B.1. Guideline-Based Schema Construction

The objective of this stage is to distill a fixed, principled schema of determinative attributes from authoritative sources. We first aggregate relevant clinical guidelines (e.g., ESTRO ACROP, NCCN, RTOG) retrieved from medical databases such as PubMed. Based on this compiled corpus, we employ a Knowledge Distillation Prompt $\mathcal{P}_{\text{schema}}(\cdot)$ that instructs the LLM to act as a domain expert to synthesize a standardized attribute list. The prompt is specifically designed to identify pathological factors that dictate boundary modifications for CTV segmentation, consolidating diverse guideline terminologies into a unified schema.

Prompt 1: Guideline Knowledge Distillation

System Role: You are a board-certified radiation oncologist and expert in prostate cancer radiotherapy planning.
Context: Accurate delineation of the Clinical Target Volume (CTV) for postoperative prostate cancer relies on specific pathological risk factors defined in consensus guidelines.
Task: Read the aggregated guideline documents provided below. Identify and summarize the specific pathological attributes that explicitly govern the anatomical boundaries of the CTV. For each attribute, explain how it influences the target volume (e.g., "inclusion of seminal vesicle bed").
Input Guidelines: [Insert full text of compiled ESTRO / NCCN / RTOG guidelines here...]
Requirements:

1. Output a structured list of determinative attributes (e.g., T-Stage, Gleason Score).
2. Focus strictly on factors influencing anatomical target boundaries.
3. Consolidate synonymous terms into a standardized schema key.

Output Format: JSON list of keys.

Based on the agent's output, we finalized the attribute schema $\mathcal{S}$ by retaining factors with explicit spatial implications for CTV delineation. This selection was further verified by senior radiation oncologists to ensure alignment with clinical consensus. The six determinative attributes are: **Pathological T-Stage**, **Gleason Score**, **Seminal Vesicle Invasion**, **Extraprostatic Extension**, **Surgical Margin Status**, and **Lymph Node Status**.

B.2. Multi-Stage Attribute Extraction

This stage transforms lengthy, unstructured pathology reports into the structured attribute profile $\mathcal{A}$. The process involves a context retrieval step followed by semantic verification and value standardization.

1. Relevant Context Retrieval: To efficiently narrow the search space within lengthy patient records, we employ a Context Retrieval Prompt that functions as the operator $\mathcal{F}_{\text{key}}(\cdot)$. This step filters the raw document $\mathbf{D}_{\text{raw}}$ to identify candidate text spans related to the schema $\mathcal{S}$, strictly excluding irrelevant medical history. The output list constitutes the **relevant text set $\mathbf{T}_{\text{rel}}$**, which serves as the input for the subsequent verification step.

Prompt 2: Relevant Context Retrieval

System Role: You are assisting in extracting pathological attributes for postoperative prostate cancer.

Task: Given the schema below, identify all sentences or short text spans from the pathology report that may contain information relevant to any attribute in the schema.

Schema: [Insert Attribute Schema $\mathcal{S}$ derived from Prompt 1]

Pathology Report: [Insert Raw Pathology Report $\mathbf{D}_{\text{raw}}$]

Output:

1. A list of relevant text spans (verbatim from the report).
2. Do NOT infer values yet; only retrieve candidate segments.

---

2. Semantic Verification & Standardization: We design a Clinical Extraction Prompt to process the retrieved context $\mathbf{T}_{\text{rel}}$. This prompt forces the agent to verify the semantic context (e.g., distinguishing "margins are negative" from "margins were not assessed") and standardize the values for downstream reasoning.

Prompt 3: Attribute Verification and Standardization

System Role: You are an expert pathologist. Your task is to extract structured clinical variables from the provided text segments of a radical prostatectomy pathology report.

Input Text: [Insert filtered text segments $\mathbf{T}_{\text{rel}}$ from Prompt 2]

Target Schema: Extract values for the following attributes:

1. Pathological T-Stage
2. Gleason Score (e.g., 7(3+4))
3. Seminal Vesicle Invasion (SVI)
4. Extraprostatic Extension (EPE)
5. Surgical Margin Status
6. Lymph Node Status

Instructions:

1. **Semantic Verification:** Ignore text related to previous biopsy history or other irrelevant procedures. Focus only on the final surgical pathology.
2. **Redundancy Removal:** If multiple mentions exist, prioritize the "Final Diagnosis" section.
3. **Standardization:** Map the extracted values to the following standard formats:
   - SVI/EPE/Margins/Nodes: "Positive" or "Negative".
   - T-Stage: e.g., "pT2", "pT3a", "pT3b".
   - If an attribute is not mentioned or cannot be determined, output "Unknown".

Output Format: Provide the result as a JSON object: {"Attribute": "Standardized Value"}.

B.3. Additional Experiments

B.3.1. Qualitative Comparison

Figure 4 presents a qualitative comparison of segmentation results across five representative patients from the test set. Overall, ReaCT produces contours that exhibit superior anatomical consistency and adherence to clinical guidelines compared to competing methods.
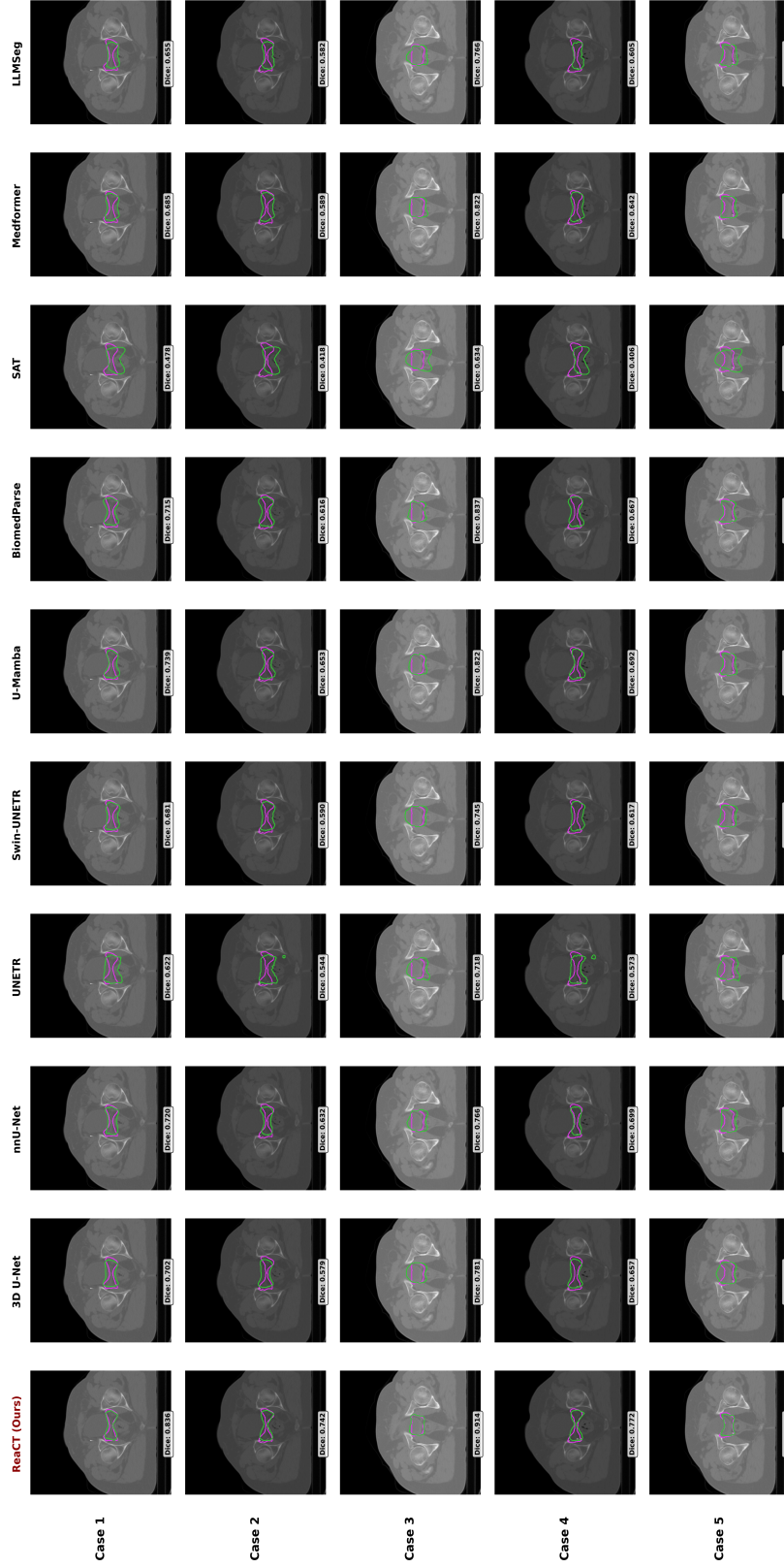
Figure 4: Qualitative comparison on five representative cases. The Ground Truth is outlined in pink, and model predictions are in green. Columns: **ReaCT (Ours)**, 3D U-Net, nnU-Net, BiomedParse, and LLMSeg. ReaCT consistently demonstrates superior alignment with the ground truth in radiographically ambiguous regions compared to baseline methods.

Table 2: Robustness Analysis. Performance degradation when individual clinical attributes are corrupted. Each row shows results when a single attribute is deliberately flipped to its opposite clinical value (e.g., Positive ↔ Negative) while keeping all other attributes correct.

Model Variant	Dice ↑	HD95 (mm) ↓	ASSD (mm) ↓
ReaCT (Original)	$\mathbf{0.8185}_{\pm 0.05}$	$\mathbf{4.15}_{\pm 1.66}$	$\mathbf{1.38}_{\pm 0.48}$
ReaCT w/ Wrong SM	$0.7924_{\pm 0.06}$ (-2.61%)	$4.82_{\pm 1.89}$ $(+0.67)$	$1.56_{\pm 0.52}$ $(+0.18)$
ReaCT w/ Wrong EPE	$0.7979_{\pm 0.05}$ (-2.06%)	$4.68_{\pm 1.78}$ $(+0.53)$	$1.52_{\pm 0.51}$ $(+0.14)$
ReaCT w/ Wrong Stage	$0.7995_{\pm 0.05}$ (-1.90%)	$4.59_{\pm 1.74}$ $(+0.44)$	$1.49_{\pm 0.50}$ $(+0.11)$
ReaCT w/ Wrong GS	$0.8036_{\pm 0.05}$ (-1.49%)	$4.42_{\pm 1.71}$ $(+0.27)$	$1.45_{\pm 0.49}$ $(+0.07)$
ReaCT w/ Wrong LNS	$0.8064_{\pm 0.05}$ (-1.21%)	$4.35_{\pm 1.69}$ $(+0.20)$	$1.43_{\pm 0.49}$ $(+0.05)$
ReaCT w/ Wrong SVI	$0.8077_{\pm 0.05}$ (-1.08%)	$4.31_{\pm 1.68}$ $(+0.16)$	$1.42_{\pm 0.49}$ $(+0.04)$

B.3.2. Robustness and Sensitivity Analysis

To verify that ReaCT actively leverages clinical attributes for decision-making rather than treating text as a passive feature, we conducted a robustness analysis by deliberately corrupting individual attributes during inference. Specifically, for each experiment, we flipped the value of a single determinative attribute to its opposite clinical status (e.g., changing *Surgical Margin Status* from "Positive" to "Negative" or vice versa) while keeping all other attributes and the image input unchanged. This setup isolates the causal impact of each specific attribute on the segmentation outcome. As shown in Table 2, incorrect pathological inputs consistently degrade performance. Notably, corrupting the Surgical Margin status causes the most significant drop (Dice -2.61%, HD95 $+0.67$ mm). This aligns with clinical guidelines where positive margins mandate aggressive CTV expansion, significantly altering target geometry. Similarly, incorrect Extraprostatic Extension and Pathological Stage inputs also lead to marked performance losses, confirming the model's dependency on accurate determinative factors.