PTCOG-AO2025-ABS-0091

Vision-language Model for Enhanced Organ Segmentation in Proton/Heavy Ion Radiotherapy

Chenbin Liu[*,1], Yihao Zhao[1], Cuiyun Yuan[1], Ying Liang[1], Yang Li[1], Chunxia Li[1], Man Zhao[1], Jun Hu[1], Ningze Zhong[1], Wei Liu[2]

[*1] Radiation Oncology, Chinese Academy of Medical Sciences, Cancer Hospital, Shenzhen Center, China, [2] Radiation Oncology, Mayo Clinic Arizona, United States of America

Objectives

Accurate delineation of organs-at-risk is critical for proton and heavy ion radiotherapy planning, as precise dose delivery depends heavily on robust segmentation of anatomical structures. Current automated segmentation methods often underutilize prior anatomical knowledge and expert textual descriptions, limiting their accuracy and clinical applicability. This study aims to develop and validate a vision-language model that leverages detailed organ descriptions to improve segmentation performance specifically for proton/heavy ion therapy workflows.

Methods

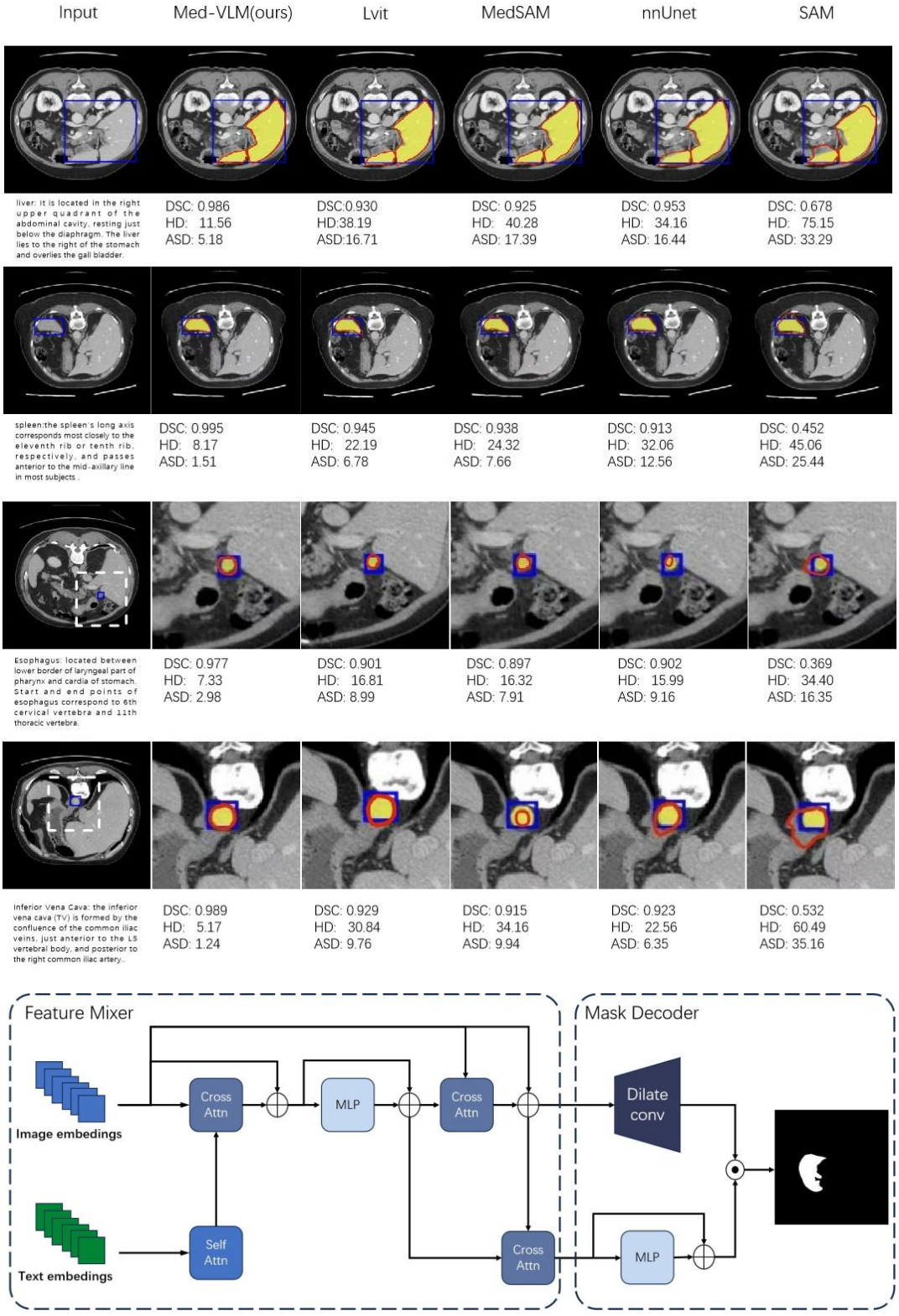
We develop Med-VLM, a novel medical image segmentation framework that integrates transformer-based vision encoders (ViT-base) with text encoders (BioBERT) using low-rank adaptation (LoRA) and a query-based feature mixer. The model utilizes expert-provided descriptions of organ location and spatial relationships as input, alongside CT image slices. During training, image and text encoder weights remain frozen, while adapters, the feature mixer, and a mask decoder are fine-tuned. The mask decoder employs dilated convolutions and a bi-directional transformer to generate precise segmentation masks. Training was conducted on three large medical image datasets (FLARE, SegTHOR, and MSD) comprising 47,000 images and 100,000 high-quality masks. Segmentation performance was evaluated using the Dice Similarity Coefficient (DSC), 95th percentile Hausdorff Distance (HD95), and Average Surface Distance (ASD).

Results

Our model demonstrated superior segmentation accuracy compared to state-of-the-art methods (nnUnet, SAM, MedSAM, Lvit) across all datasets. On the combined test set, Med-VLM achieved DSC 0.98±0.03 (vs. 0.95±0.03 for nnUnet), HD95 21.86±39.95 (vs. 45.13±62.61 for nnUnet), and ASD 4.56±4.67 (vs. 13.11±13.94 for nnUnet). The incorporation of authoritative organ descriptions significantly improved boundary delineation, especially in regions critical for proton/heavy ion therapy planning. Ablation studies confirmed the importance of detailed textual input and the effectiveness of LoRA and BioBERT-based text encoding.

Conclusions

Our model represents a significant advance in automatic organ segmentation for proton and heavy ion radiotherapy, leveraging expert textual descriptions to achieve state-of-the-art accuracy and boundary delineation. This approach has strong potential to streamline and enhance clinical workflows in advanced radiotherapy, supporting more precise and reliable dose delivery to target volumes while minimizing exposure to organs-at-risk.


Input
Med-VLM(ours)
Lvit
MedSAM
nnUnet
SAM
liver: it is located in the right upper quadrant of the abdominal cavity, resting just below the diaphragm. The liver lies to the right of the stomach and overlies the gall bladder.
DSC: 0.986
HD: 11.56
ASD: 5.18
DSC:0.930
HD:38.19
ASD:16.71
DSC: 0.925
HD: 40.28
ASD: 17.39
DSC: 0.953
HD: 34.16
ASD: 16.44
DSC: 0.678
HD: 75.15
ASD: 33.29
spleen:the spleen's long axis corresponds most closely to the eleventh rib or tenth rib, respectively, and passes anterior to the mid-axillary line in most subjects .
DSC: 0.995
HD: 8.17
ASD: 1.51
DSC: 0.945
HD: 22.19
ASD: 6.78
DSC: 0.938
HD: 24.32
ASD: 7.66
DSC: 0.913
HD: 32.06
ASD: 12.56
DSC: 0.452
HD: 45.06
ASD: 25.44
Esophagus: located between lower border of laryngeal part of pharynx and cardia of stomach. Start and end points of esophagus correspond to 6th cervical vertebra and 11th thoracic vertebra.
DSC: 0.977
HD: 7.33
ASD: 2.98
DSC: 0.901
HD: 16.81
ASD: 8.99
DSC: 0.897
HD: 16.32
ASD: 7.91
DSC: 0.902
HD: 15.99
ASD: 9.16
DSC: 0.369
HD: 34.40
ASD: 16.35
Inferior Vena Cava: the inferior vena cava (IV) is formed by the confluence of the common iliac veins, just anterior to the L5 vertebral body, and posterior to the right common iliac artery..
DSC: 0.989
HD: 5.17
ASD: 1.24
DSC: 0.929
HD: 30.84
ASD: 9.76
DSC: 0.915
HD: 34.16
ASD: 9.94
DSC: 0.923
HD: 22.56
ASD: 6.35
DSC: 0.532
HD: 60.49
ASD: 35.16
Feature Mixer
Image embedings
Text embedings
Cross Attn
MLP
Cross Attn
Self Attn
Cross Attn
Mask Decoder
Dilate conv
MLP