



ConTEXTual Net CT: A 3D Visual Grounding Model for Abdominal CT

Samuel D. Church, MS, University of Wisconsin – Madison; Joshua D. Warner, MD, PhD;
Danyal Maqbool, MS; Junjie Hu, PhD; Meghan G. Lubner, MD; Tyler J. Bradshaw, PhD

Introduction/Background

Medical visual grounding models localize image regions corresponding to phrases in clinical reports, offering potential for enhanced disease tracking and reporting. In 3D domains like CT, progress is limited by the cost of large-scale annotated datasets. We present ConTEXTual Net CT, a visual grounding model trained using an automated labeling pipeline that converts existing PACS annotations into image-phrases-mask training triplets.

Methods/Intervention

Our pipeline extracts radiologist-drawn GSPS annotations (e.g., lines, arrows) from PACS, converts them to 3D segmentations via the promptable segmentation model SAM2CT, and pairs the resulting masks with corresponding report phrases using shared metadata (e.g., measurements, series/slice numbers). This generated 105,130 weakly-labeled image-phrases-mask triplets from 51,567 abdominal CT exams for training. Our model, ConTEXTual Net CT, is a UNet-based architecture that fuses text and image features via cross-attention in the decoder, using a frozen Qwen3-Embedding-8B as the text encoder. Images were windowed to standard lung, bone, and soft-tissue windows, and stacked as multi-channel inputs. We evaluated visual grounding performance on a test set of 127 emergency department (ED) cases spanning 13 common categories and 98 oncology cases, with labels generated via our pipeline and manually refined until approved by board-certified radiologists. Performance was compared against BiomedParse and VoxTell, the current state-of-the-art CT visual grounding models.

Results/Outcome

ConTEXTual Net CT outperformed BiomedParse and VoxTell on all 13 ED categories, achieving an average Dice of 0.427 and hit rate (HR=proportion of cases with Dice ≥ 0.1) of 0.720. Performance was strongest on hernias (HR=0.9). The best comparator model's average Dice was 0.132, with a HR of 0.248. ConTEXTual Net CT also outperformed comparators on the oncology test set, achieving an average Dice score of 0.455 and HR of 0.66, while the best comparator averaged a Dice of 0.136 and HR of 0.213.

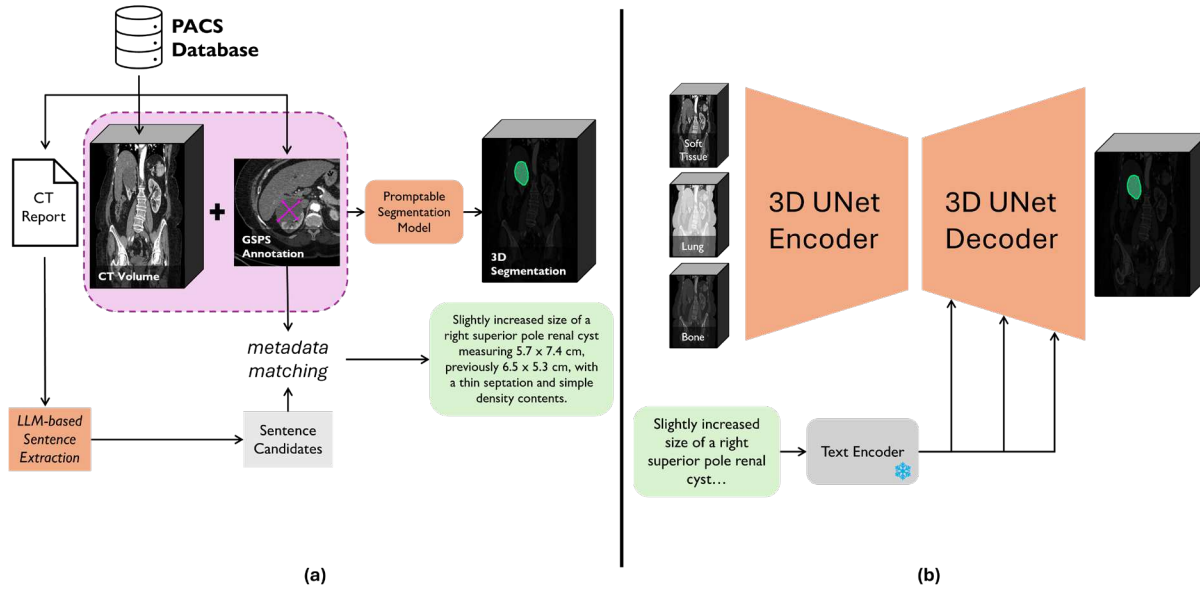
Conclusion

We developed ConTEXTual Net CT, a visual grounding model for abdominal CT, which outperformed comparator models. Our automated labeling pipeline enables effective training of 3D visual grounding models on diverse CT findings, without costly manual annotation.

Statement of Impact

By generating large-scale weakly-labeled visual grounding datasets from existing PACS annotations, our

pipeline removes a bottleneck to training visual grounding models, allowing for enhanced reporting.



Overview of the data labeling pipeline and model architecture. a) Data labeling pipeline. CT volumes, free-text reports, and associated GSPS annotations (lines, arrows) are extracted from the PACS database. GSPS annotations are converted to 3D segmentations using the promptable segmentation model SAM2CT. An LLM extracts candidate sentences from the report and matches each to a GSPS annotation based on measurement or series/slice number, yielding a 3D segmentation paired with a referring expression from the original report. b) ConTEXTual Net CT architecture. Three windowed versions of the image (soft tissue, lung, bone) are encoded by a 3D UNet; the referring expression is encoded by a text encoder. Text and image features are fused via cross-attention at each stage of the UNet decoder, producing a predicted segmentation corresponding to the referring expression.

Category	ConTEXTual Net CT		VoxTell		BiomedParse	
	Dice	Hit Rate	Dice	Hit Rate	Dice	Hit Rate
Emergency Department (N=125)	0.427	0.720	0.132	0.248	0.071	0.200
Abscess (10)	0.244	0.600	0.000	0.000	0.001	0.000
Aneurysm (10)	0.542	0.800	0.188	0.400	0.222	0.700
Appendicitis (9)	0.339	0.667	0.000	0.000	0.000	0.000
Cyst (10)	0.520	0.800	0.116	0.200	0.096	0.200
Diverticulitis (9)	0.106	0.222	0.000	0.000	0.000	0.000
Gallstone (10)	0.517	0.800	0.375	0.600	0.080	0.300
Hematoma (8)	0.416	0.750	0.040	0.125	0.109	0.250
Hernia (10)	0.659	0.900	0.000	0.000	0.018	0.100
Kidney Stone (10)	0.334	0.700	0.270	0.500	0.004	0.000
Lesion (10)	0.482	0.800	0.195	0.500	0.036	0.100
Lymph Node (10)	0.487	0.800	0.024	0.100	0.093	0.400
Mass (9)	0.473	0.777	0.227	0.333	0.135	0.222
Nodule (10)	0.392	0.700	0.244	0.400	0.126	0.300
Oncology (N=97)	0.455	0.660	0.136	0.213	0.075	0.160

3D Dice scores and hit rate (HR=proportion of cases with Dice $\geq$ 0.1) on both the emergency department and oncology test sets.

Emergency Department

Peripherally calcified 1.5 cm splenic artery aneurysm is unchanged compared to [DATE].



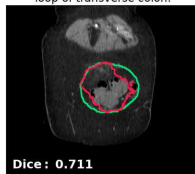
Moderate uncomplicated diverticulitis with short segment of mural thickening proximal sigmoid, with pericolic stranding most prominent along a diverticula (series 2/93).



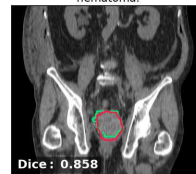
Dilated hyperemic appendix to 9 mm, with surrounding fat stranding in the left lower quadrant.



Midline abdominal parastomal hernia measuring 16.6 x 9.2 x 12.2 cm, now containing a loop of transverse colon.



Pelvis: Large soft tissue density within the pelvis at the expected location of the prostate measuring 4.8 x 7.0 x 7.2 cm, which is concerning for hematoma.



Oncology

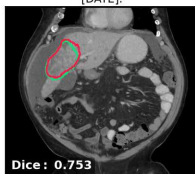
The fat density mass along the lesser curvature of the stomach is similar, measuring 4.4 x 7.3 cm compared to 5.3 x 6.7 cm on [DATE].



Increased size of a left periaortic retroperitoneal lymph node (series 2 image 46) measuring up to 1.0 cm.



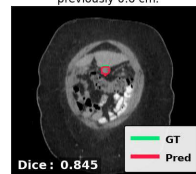
Continued slight increase in size of the dominant infiltrative mass in segment 4/8 of the liver which now measures approximately 7.1 x 12.0 x 8.3 cm as compared to 6.6 x 11.2 x 7.9 cm when measured similarly on [DATE].



1.8 cm nodule in the left upper quadrant with an associated calcification (series 2, image 52), stable from the prior study, and likely representing fat necrosis.



Increased size of a soft tissue nodule within the central anterior abdomen just inferior to the left hepatic lobe, now measuring up to 2.1 cm, previously 0.6 cm.



Example ConTEXTual Net CT outputs on emergency department and oncology test cases. Predicted and ground truth masks are shown in red and green, respectively.

Keywords

Deep Learning; CT; Automated Labeling; Visual Grounding