GroundingDINO for Open-Set Lesion Detection in Medical Imaging

Samuel J. Roughley[1], Johanna P. Müller[2], Shangqi Gao[3,4], Zeyu Gao[3,4], Marta Ligero[5], Rudolfs Blums[6], Mireia Crispin-Ortuzar[3,4], Julia Schnabel[6,7,8,9], Bernhard Kainz[2,10], Cosmin I. Bercea[6,7], and Ines Prata Machado[3,4]

[1] Department of Physics, University of Cambridge [2] Friedrich-Alexander University Erlangen-Nürnberg [3] Department of Oncology, University of Cambridge [4] Early Cancer Institute, University of Cambridge [5] Else Kroener Fresenius Center for Digital Health, Technical University Dresden [6] Technical University of Munich [7] Helmholtz AI and Helmholtz Center Munich [8] Munich Center for Machine Learning [9] King's College London [10] Imperial College London

Motivation

🔎 Radiological error rate of ~1/3[1,2]

💻 Computational diagnostic tools must demonstrate robust performance on rare pathologies

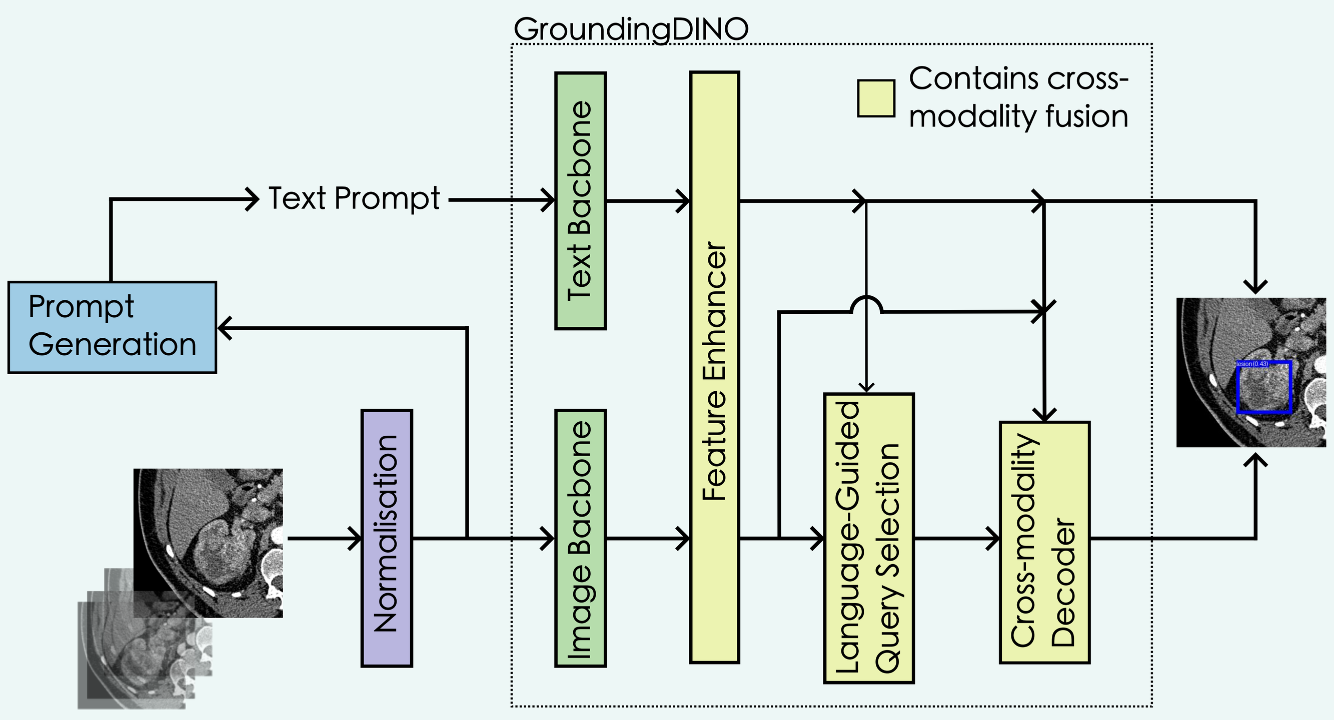
🧠 Natural imaging methods incorporate semantic information to obtain open-set capabilities

Can we use language to improve the detection of lesions on CT scans?

Contributions:

1. First investigation of GroundingDINO model[3] for medical anomaly detection
2. Examine impact of prompt design on closed-set and open-set performance

Method



Modular design allows for multiple points of **cross-modality fusion**

Provide semantic information through a text prompt to guide the detection, enabling **open-set** capabilities

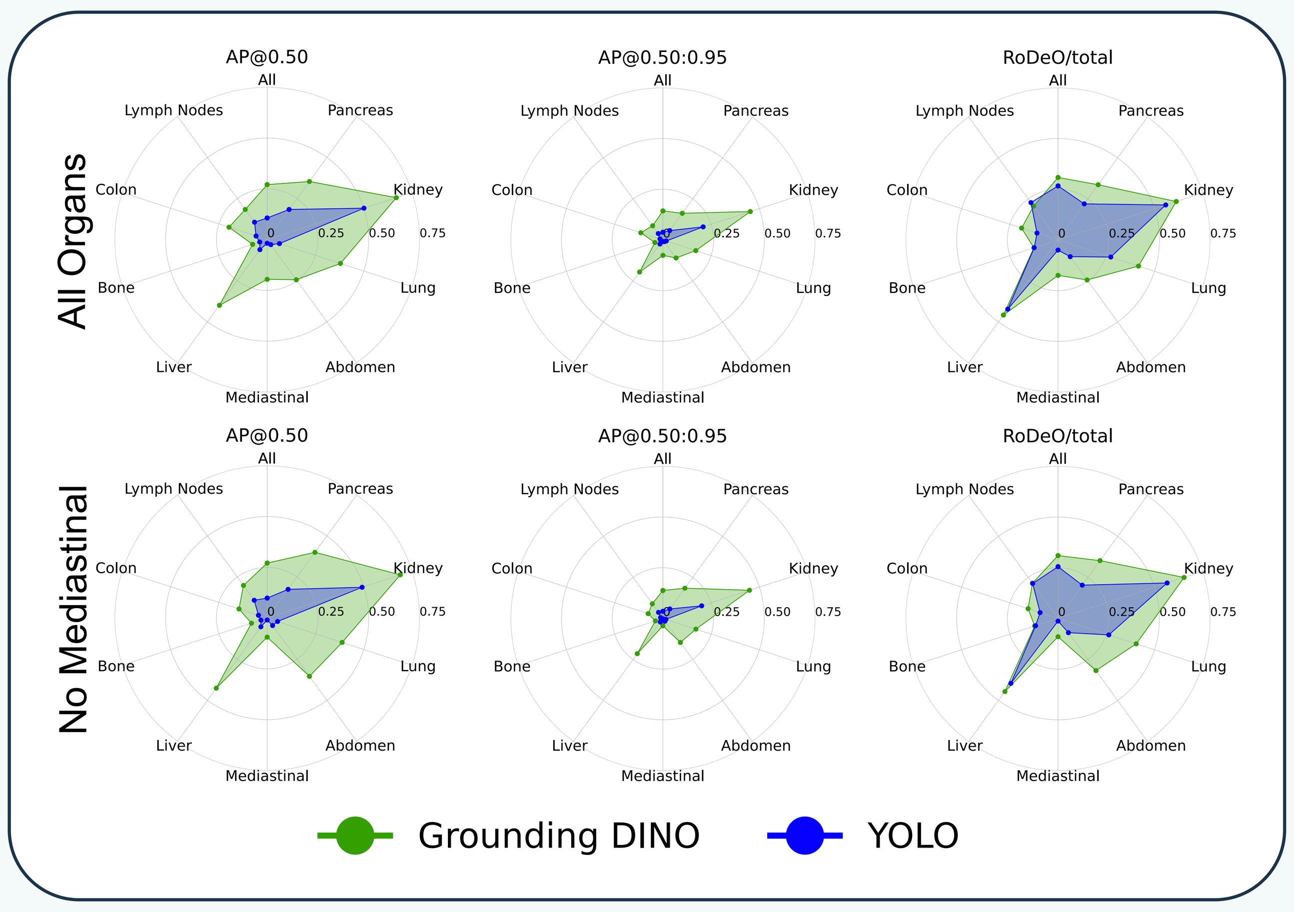
Three types of text prompt explored:

I. Generic (*"lesion"*)
II. Organ-specific (*"[organ] lesion"*)
III. Visual description

Increasing semantic information

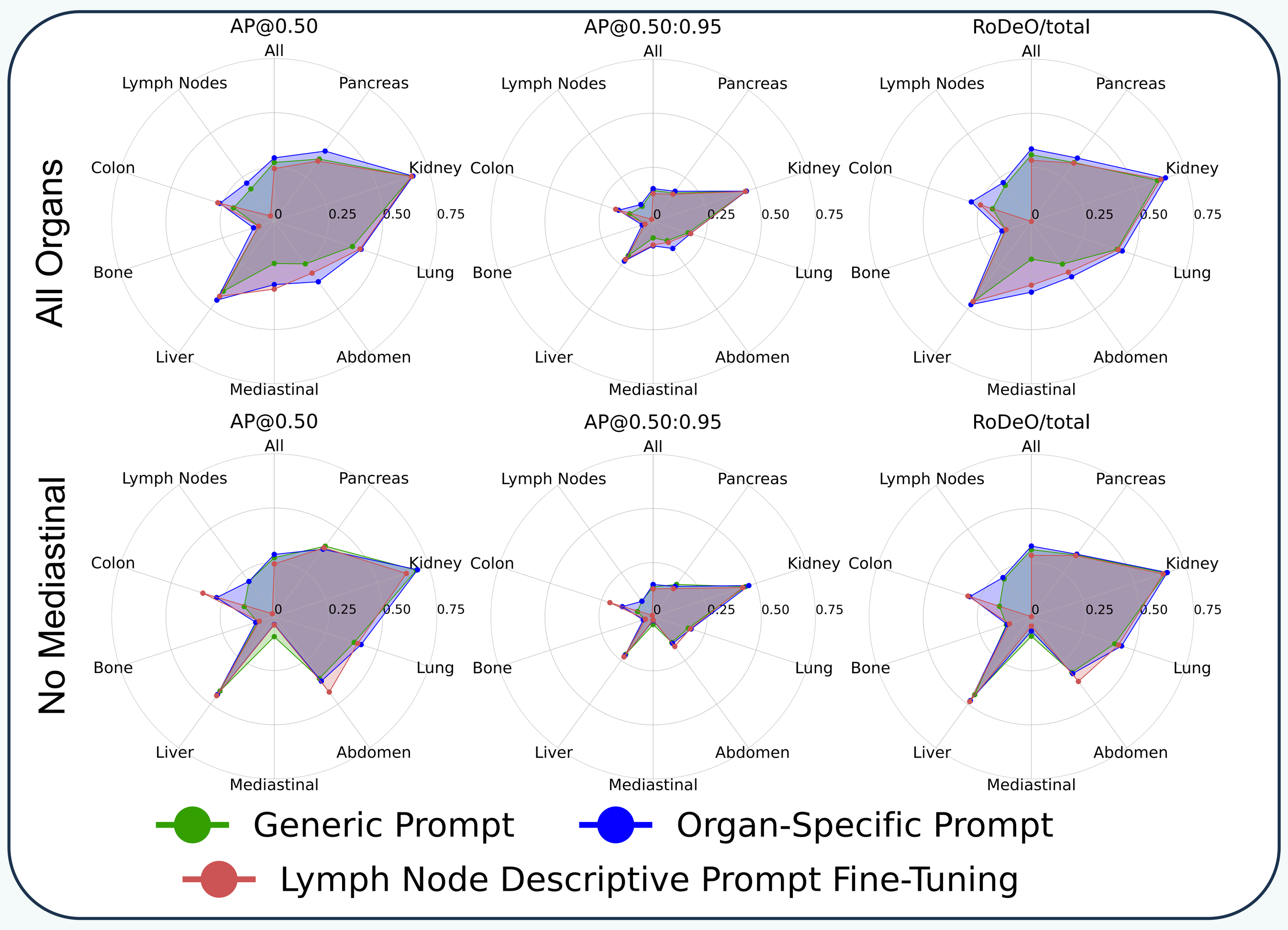
Dataset: CT scans of **chest-abdomen-pelvis** region, containing 6,382 lesions across range of organs

Results

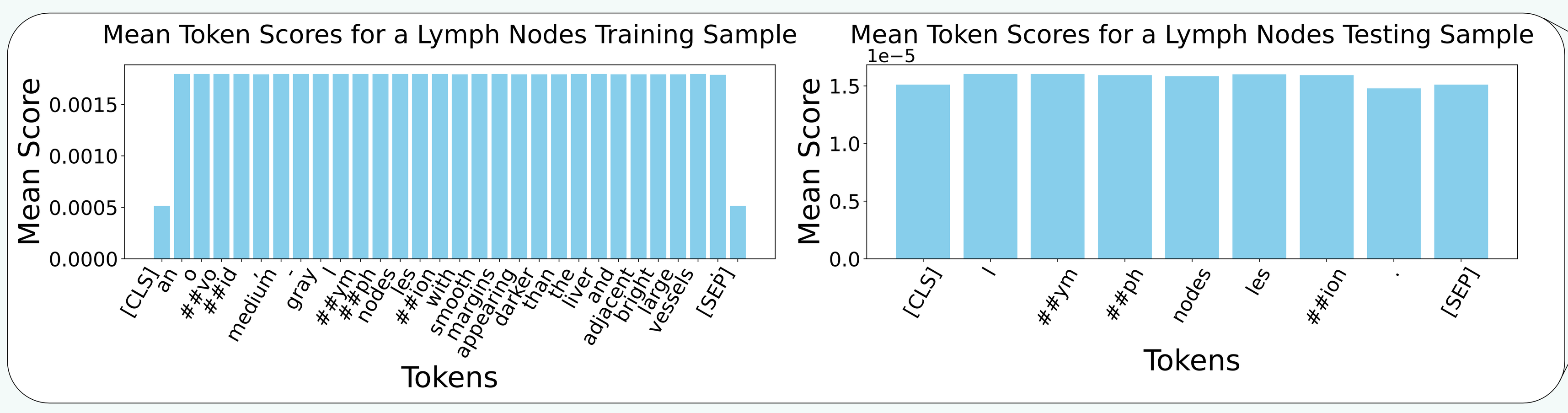


GroundingDINO **outperforms YOLOv11n** on closed- and open-set performance with minimal contextual information (generic prompts)

Enhanced language guidance



Organ specification improves closed-set performance on rarer lesion classes



Introducing visual descriptions highlights **overfitting**; GroundingDINO memorises prompt-image pairs, highlighted by **uniform activation** across prompts

Future work: Changes to loss function, text encoder, or further prompt engineering may be required to better leverage language-based cues

References:

1. Kim & Mansfield., AJR, 2014, 10.2214/ajr.13.11493
2. Berlin, AJR, 2007, 10.2214/ajr.06.1270
3. Liu et al. 2024, *arXiv:2303.05499*

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

SCAN TO DISCOVER MORE

Acknowledgments. C.I.B. is funded via the EVUK programme ("Next-generation AI for Integrated Diagnostics") of the Free State of Bavaria and partially supported by the Helmholtz Association [Munich School for Data Science]. This work is also supported by the Berdelle-Stiftung [TimeFlow]. The authors acknowledge scientific support and HPC resources from NHR@FAU [b143dc, b180dc], funded by federal and Bavarian authorities, with partial hardware funding from the DFG [440719683]. Additional support was received from the ERC [101083647], the DFG [KA 5801/2-1, INST 90/1351-1], and the state of Bavaria. Further funding was provided by Cancer Research UK [A22905], the CRUK Cambridge Centre [CTRQQR-2021-100012, A25177], The Mark Foundation for Cancer Research [RG95043], GE HealthCare, the CRUK National Cancer Imaging Translational Accelerator [A27066], and the NIHR Cambridge Biomedical Research Centre [NIHR203312, BRC-1215-20014].