SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

A 3D Merlin-Based Vision-Language Model for Multi-Injury Detection in Trauma CT Using Report-Grounded Supervision

Mehrdad Salimitari, PhD, University of Maryland Baltimore; Taiyu Zhang, PhD; Cameron May, MS; Guang Li, MD; David Dreizin, MD

Introduction/Background
Whole-body trauma CT interpretation is complex and time-critical, requiring rapid detection of a broad spectrum of injuries with highly variable prevalence. Existing AI approaches typically focus on single injuries or require extensive manual labeling, limiting scalability across heterogeneous trauma presentations. We developed a vision-language model (VLM) for automated detection of diverse traumatic injuries from CT using paired radiology-report supervision.

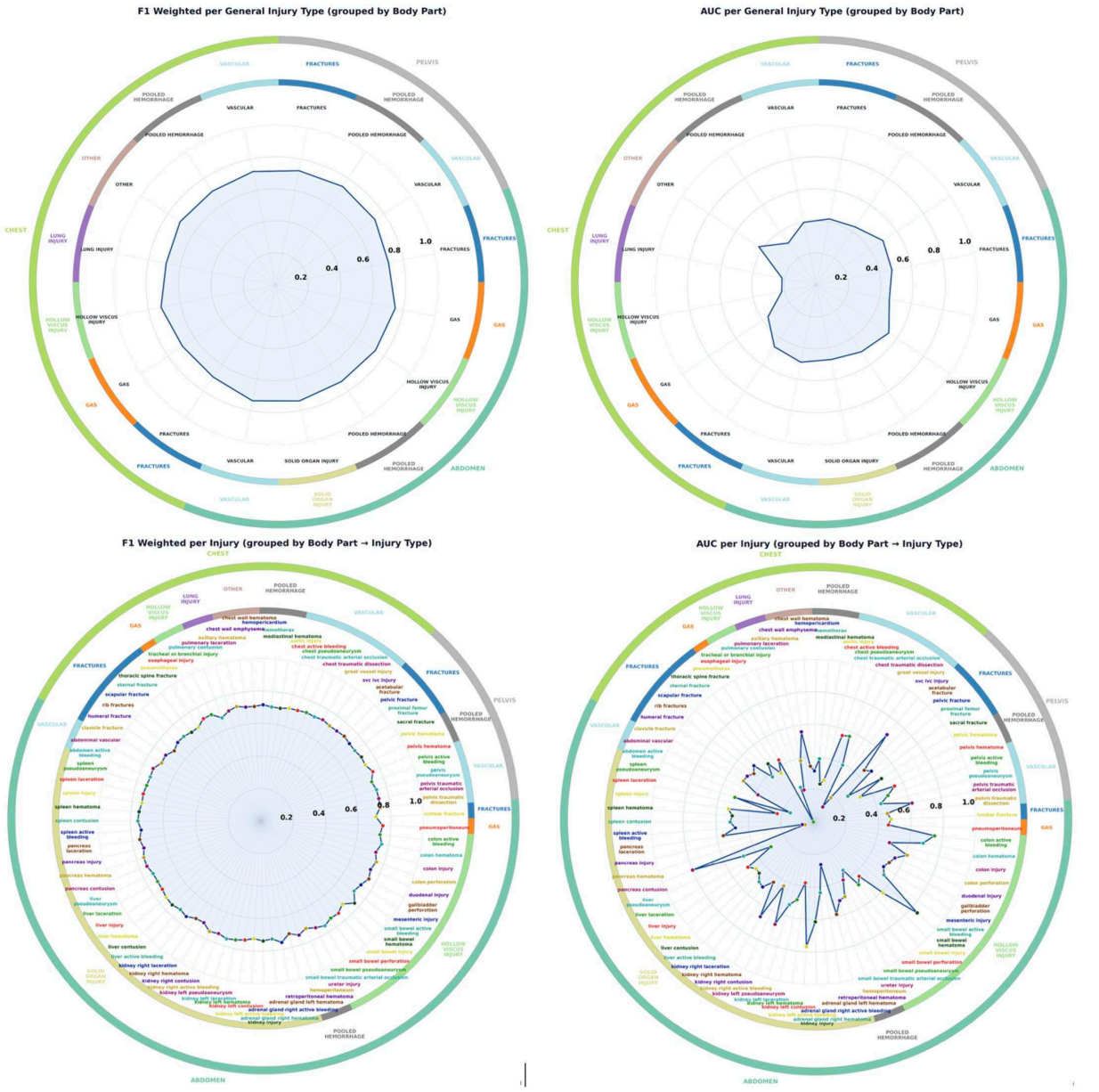
Methods/Intervention
We included 8,700 consecutive trauma CT examinations from our trauma center (60/40 train/validation split), encompassing 83 distinct injuries across the chest, abdomen, and pelvis with highly imbalanced prevalence. Positive injury labels were extracted from radiology reports using the GPT-5-mini model. Because absent findings are typically not explicitly documented, negative injury labels were synthetically generated using templated, anatomy-aware report text to balance positive and negative supervision. A 3D Merlin-based vision-language model was trained on paired CT images and text labels using contrastive learning on an 8×H100 server. CT preprocessing used injury-aware windowing (bone, soft tissue, and lung) and injury-specific phase selection (arterial, portal venous, or both). Performance was evaluated per injury using area under the ROC curve (AUC) and weighted F1 score, with additional grouping by body region and injury category.

Results/Outcome
Across all 83 injuries, the weighted F1 score was 0.73 (per-injury range ≈0.6–0.8) despite substantial class imbalance. AUC values for most injuries ranged between 0.5 and 0.8, with higher performance observed for common and imaging-salient injuries and lower AUC for rare or subtle findings. Grouped analysis demonstrated stable performance across major body regions and injury types, supporting generalizability across heterogeneous trauma patterns.

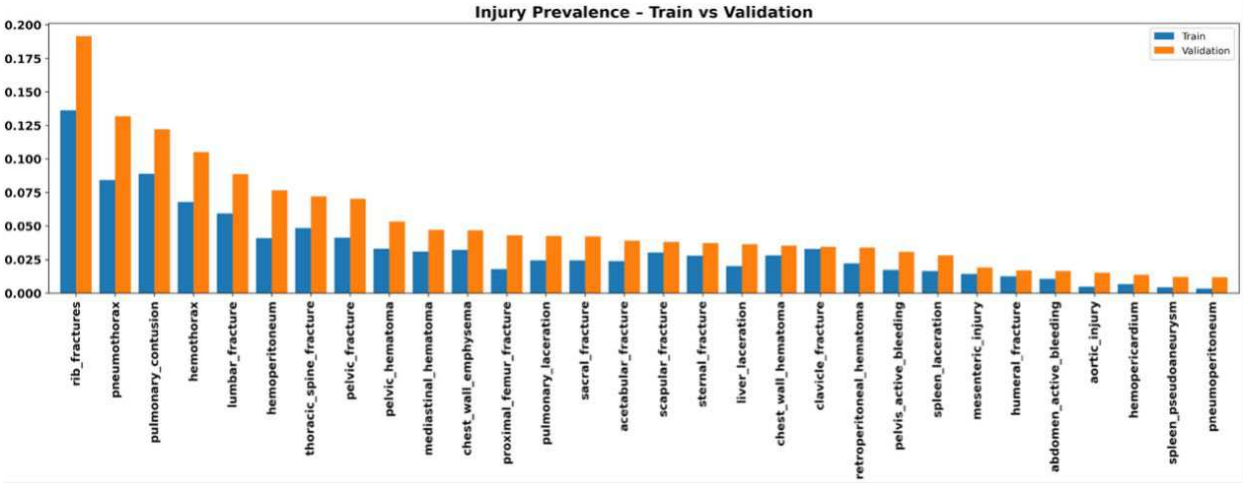
Conclusion
A report-supervised vision-language model can detect a broad range of traumatic injuries from CT without manual annotation, establishing an initial Merlin-derived baseline (weighted F1 0.73) for comprehensive trauma detection. Synthetic negative generation further enables robust training from routinely available clinical reports.

Statement of Impact
Report-supervised training removes the manual-annotation bottleneck confining most trauma AI to single injuries. By deriving labels from routine radiology reports and pairing them with synthetic negatives, one model

spans 83 injuries across four body regions—a scalable, low-cost route toward comprehensive whole-body triage tools that could speed CT interpretation and reduce missed injuries in trauma care.



Spider plots show weighted F1s and support-weighted mean AUCs grouped by body region and broad injury class (top), and individual injury weighted F1s and AUCs (bottom).



Per-injury prevalence for the 30 most common injuries in train and validation sets demonstrate the long tail encountered in trauma whole-body CT.

Keywords

Trauma CT; Vision-Language Models; Artificial Intelligence; Trauma Detection; Radiology Informatics