npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01807-8

Vision-language model for report generation and outcome prediction in CT pulmonary angiogram

Zhusi Zhong[1,2], Yuli Wang[3], Jing Wu[4], Wen-Chi Hsu[5,6], Vin Somasundaram[1,2], Lulu Bi[1,2], Shreyas Kulkarni[1,2], Zhuoqi Ma[1,2], Scott Collins[1,2], Grayson Baird[1,2], Sun Ho Ahn[1,2], Xue Feng[7], Ihab Kamel[8], Cheng Ting Lin[6], Colin Greineder[9], Michael Atalay[1,2], Zhicheng Jiao[1,2] ✉ & Harrison Bai[6,8] ✉

Accurate and comprehensive interpretation of pulmonary embolism (PE) from Computed Tomography Pulmonary Angiography (CTPA) scans remains a clinical challenge due to the limited specificity and structure of existing AI tools. We propose an agent-based framework that integrates Vision-Language Models (VLMs) for detecting 32 PE-related abnormalities and Large Language Models (LLMs) for structured report generation. Trained on over 69,000 CTPA studies from 24,890 patients across Brown University Health (BUH), Johns Hopkins University (JHU), and the INSPECT dataset from Stanford, the model demonstrates strong performance in abnormality classification and report generation. For abnormality classification, it achieved AUROC scores of 0.788 (BUH), 0.754 (INSPECT), and 0.710 (JHU), with corresponding BERT-F1 scores of 0.891, 0.829, and 0.842. The abnormality-guided reporting strategy consistently outperformed the organ-based and holistic captioning baselines. For survival prediction, a multimodal fusion model that incorporates imaging, clinical variables, diagnostic outputs, and generated reports achieved concordance indices of 0.863 (BUH) and 0.731 (JHU), outperforming traditional PESI scores. This framework provides a clinically meaningful and interpretable solution for end-to-end PE diagnosis, structured reporting, and outcome prediction.

Pulmonary Embolism (PE) is a life-threatening condition caused by blood clots obstructing the pulmonary arteries, often leading to severe complications, long-term morbidity, and a high risk of mortality. In the United States alone, PE affects approximately 600,000 individuals annually and contributes to more than 60,000 deaths[1,2]. Despite advances in diagnostic technologies[3,4], timely and accurate PE diagnosis remains a significant clinical challenge[5,6]. Computed Tomography Pulmonary Angiography (CTPA)[7] is the gold standard for PE detection; however, interpretation can be delayed by radiologist availability, physician fatigue, and inherent complexity of the cases. Studies show that up to 30% of untreated PE cases result in death within one month of diagnosis, underscoring the urgent need for more efficient and reliable diagnostic solutions[8].

Timely and comprehensive diagnosis of PE is essential for improving patient outcomes. Current PE management rely strategies heavily on tools such as the Pulmonary Embolism Severity Index (PESI), which are based on a limited set of clinical variables[9]. However, these tools may fail to capture important contributors to disease severity and long-term complications, such as pulmonary hypertension and recurrence risk, which require detailed imaging and systematic reporting for accurate assessment[10].

Recent advances in vision-language models (VLMs) have demonstrated strong potential in 3D medical imaging applications, including abnormality detection, automated report generation, and clinical decision support. For instance, CT-CLIP and CT-CHAT, developed using the CT-RATE[11] dataset, introduced contrastive and chat-based frameworks for chest CT interpretation. Similarly, RadFM[12] and M3D[13] have extended foundation model capabilities across imaging modalities through large-scale multimodal datasets and instruction-tuning strategies. While these models offer generalizable solutions, their application to PE-specific diagnosis and

[1]Department of Diagnostic Imaging, Brown University Health, Providence, RI, USA. [2]Warren Alpert Medical School of Brown University, Providence, RI, USA. [3]Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [4]Second Xiangya Hospital, Central South University, Changsha, Hunan, China. [5]Department of Medical Imaging and Intervention, Chang Gung Memorial Hospital at Linkou, Taoyuan, Taiwan, ROC. [6]Department of Radiology and Radiological Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [7]Carina AI, Lexington, KY, USA. [8]Department of Radiology, University of Colorado School of Medicine, Aurora, CO, USA. [9]Department of Emergency Medicine and Department of Pharmacology, University of Michigan, Ann Arbor, MI, USA. ✉e-mail: zhicheng_jiao@brown.edu; hbai7@jhu.edu

prognosis using CTPA remains limited. The INSPECT[14] dataset provides a valuable foundation for multimodal PE analysis; however, few existing approaches have integrated structured reporting and survival prediction into a unified, clinically actionable pipeline.

The integration of AI into PE diagnosis and management has the potential to reduce diagnostic delays, minimize human error, and enhance the reliability of outcome prediction[15,16]. Unlike general-purpose models, task-specific AI models can address the unique challenges of PE care by automatically identifying relevant abnormalities, generating structured radiology reports, and estimating long-term prognoses from multimodal data sources, including imaging and clinical records[12,17]. By combining the interpretive strength of large language models (LLMs) with the visual analytic capabilities of VLMs, these specialized systems can assist radiologists in making faster, more accurate decisions, reducing misdiagnoses, and supporting more effective treatment strategies for PE patients[18,19].

In this study, we developed and validated an agent-based framework that integrates VLMs and LLMs to enhance the PE diagnostic workflow. We introduce a structured, abnormality-guided reporting paradigm that aligns closely with radiological practice, improving both accuracy and interpretability of generated reports. The framework is evaluated across three large-scale datasets, demonstrating strong generalizability in abnormality detection and report quality across institutions. Expert evaluation further validates the clinical relevance of the generated reports, consistently favoring our structured approach over baseline methods. Finally, we introduce a multimodal survival prediction module that integrates imaging, clinical variables, diagnostic findings, and generated reports, achieving robust and interpretable prognostic performance across cohorts.

The proposed novel agent-based framework explicitly advances beyond prior efforts by unifying three key components of the PE diagnostic and prognostic workflow: (1) fine-grained abnormality detection from CTPA, (2) structured, region-aware report generation aligned with radiologist practices, and (3) multimodal survival prediction that incorporates imaging, clinical variables, and AI-generated diagnostic content. Prior models such as CT-CHAT[11] emphasize VQA-style anomaly detection, while RadFM[12] and M3D[13] focus on free-text report generation without structured localization or prognostic integration. The INSPECT dataset[14] enables PE-related outcome modeling but lacks automated structured reporting. In contrast, our framework integrates all three tasks into a clinically interpretable and operationally cohesive pipeline, validated across multi-institutional datasets and expert assessments. This comprehensive approach bridges diagnostic interpretation with downstream risk stratification, offering a more complete, real-world solution to the challenges of PE diagnosis and management.

Results

For diagnosis and report generation, we included a total of 69,761 paired CTPA image-report studies from 24,890 patients across three datasets: Brown University Health (BUH, $n = 19{,}565$), Johns Hopkins University (JHU, $n = 1077$), and the publicly available INSPECT dataset ($n = 4248$) (Fig. 1a). For survival analysis, we identified 1012 patients with confirmed PE diagnoses (BUH: 917; JHU: 95) who had complete imaging, radiology reports, PESI clinical variables, and follow-up outcome data. Demographic and clinical characteristics of the cohorts are summarized in Table 1.

Structured CTPA diagnosis and reporting framework

As shown in Fig. 2, we introduce a structured, clinically informed CTPA diagnostic and reporting framework designed to emulate radiologists' diagnostic reasoning by systematically identifying and characterizing abnormalities. This hierarchical framework (Fig. 1b) organizes the diagnostic process by evaluating seven anatomically distinct regions and detecting 32 clinically significant abnormalities on CTPA scans. Grounded in established diagnostic standards[20,21], the framework was developed in collaboration with radiologists, emergency physicians, and pulmonologists from Brown University, Johns Hopkins University, and the University of Michigan to ensure clinical relevance, consistency, and broad applicability.

The foundation of our framework is a multi-label abnormality classification module that detects the presence or absence of the 32 predefined abnormalities from CTPA imaging data. These classification outputs serve two key purposes: they enable accurate region-level abnormality identification and provide auxiliary diagnostic signals that guide downstream report generation.

To produce structured image-based reports, we implemented a CTPA reading agent that performs region-based reporting. A medical VLM was prompted with either organ-specific or abnormality-specific queries to generate localized, clinically coherent diagnostic descriptions. This multi-stage reporting pipeline mirrors real-world radiology workflows, progressing systematically from detailed anatomical evaluations to concise, actionable diagnostic summaries.

Following the region-based analysis, a report-writing agent, implemented using the Llama 3[22] LLM, was prompted to synthesize these localized findings into comprehensive and structured reports. Specifically, the agent composes a "Study Findings" section that summarizes diagnostic observations across all evaluated regions and an "Study Impression" section that highlights the most clinically significant conclusions, particularly those related to PE.

For prognosis prediction, we integrate the multimodal data, including raw CTPA images, generated reports, abnormality classification results, and PESI clinical scores, into a multimodal survival prediction module based on time-to-event Cox regression[23]. This unified, multi-task CTPA diagnostic framework effectively combines visual, textual, and clinical information to support precise abnormality detection, standardized reporting, and clinically meaningful outcome prediction, enabling comprehensive management of PE patients.

Abnormality diagnosis: multi-abnormality classifier vs. medical VLMs

We evaluated abnormality diagnosis performance using multi-label classification metrics, including accuracy (ACC), area under the receiver operating characteristic (AUROC), sensitivity, specificity, and F1 score. Our multi-abnormality classifier was compared against several state-of-the-art (SOTA) medical VLMs: CT-CHAT (a chest CT-specific VLM)[11], RadFM (a general-purpose radiology foundation model)[12], and M3D (a 3D radiology foundational VLM)[13]. These VLMs were adapted for CTPA interpretation using visual question answering (VQA) prompts targeting the same 32 abnormalities.

As illustrated in Fig. 3, our multi-abnormality classifier consistently outperformed the VLM-based approaches across all three datasets (BUH, INSPECT, JHU). In Figure 3a, it achieved the highest AUROC on BUH (78.8%), INSPECT (75.4%), and JHU (71.0%), along with superior F1 scores of 60.9%, 58.9%, and 56.7%, respectively. Region-wise analysis (Fig. 3b) further highlights the classifier's strength, with AUROC exceeding 70% in nearly all anatomical regions and F1 scores surpassing 60% in regions such as the lungs and airways, heart, and pleura. The per-abnormality radar plot (Fig. 3c) shows the classifier's consistent superiority over other SOTA models in detecting clinically significant findings such as PE, pleural effusion, and lymphadenopathy. In the dedicated PE detection task (Fig. 3d), our model achieved the highest AUROC (74.0%) and F1 score (58.7%) on the BUH dataset, outperforming all VLM-based baselines, including the PE-specific model PENet[24]. These cross-institutional results also validate the robustness and clinical relevance of our multi-abnormality classifier and establish it as a strong foundation for downstream structured CTPA report generation.

Image report comparison: holistic vs. structured generation

We evaluated two strategies for image-based radiology report generation: (1) holistic caption-based generation, in which prompted VLMs directly generate free-text reports from entire CTPA volumes, and (2) structured region-based generation, which involves extracting region-wise findings

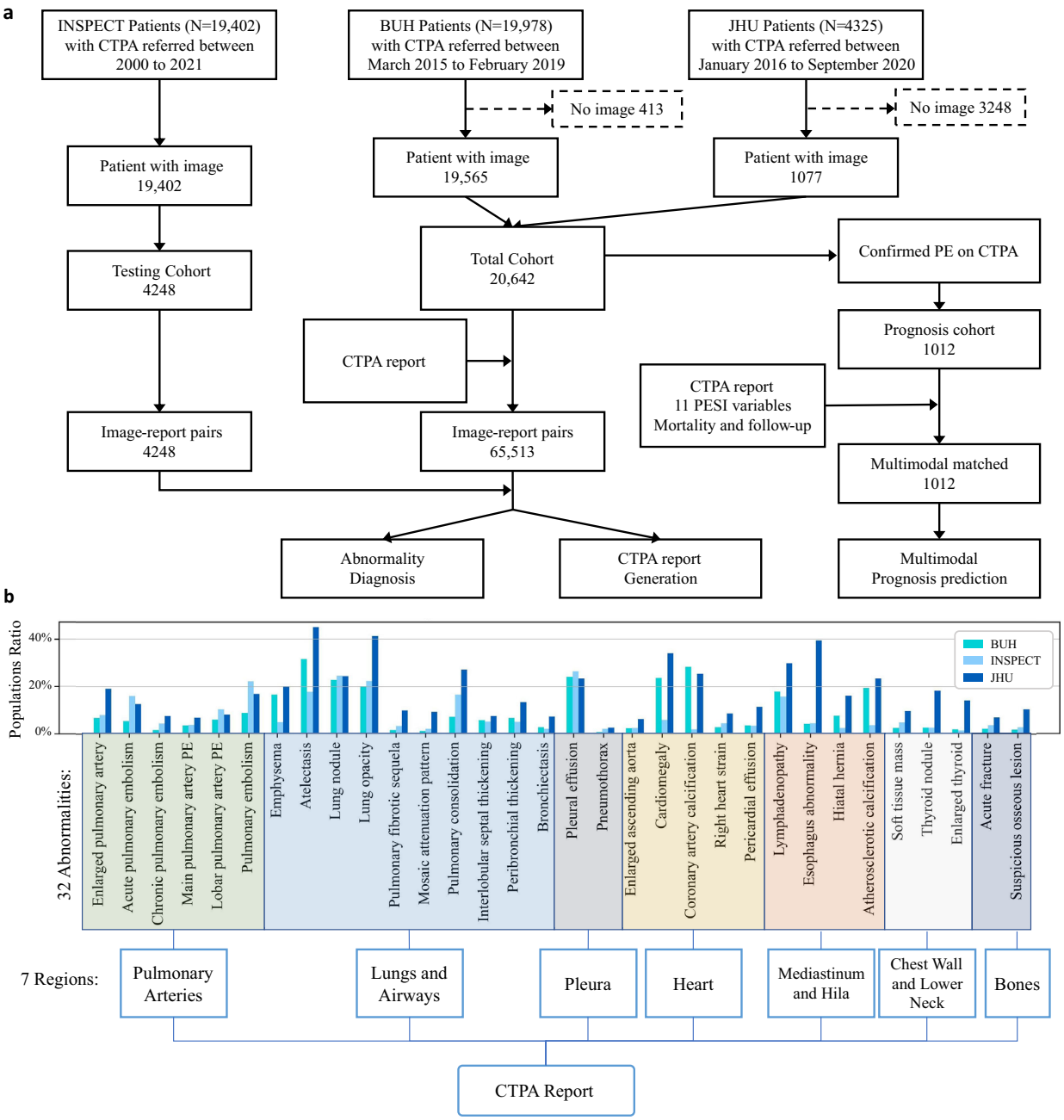


Fig. 1 | Patient flowchart. Structured diagnosis and reporting framework. **a** Flowchart illustrating the patient enrollment from three independent datasets. **b** The proposed CTPA imaging reporting framework is derived from 7 chest regions and 32 abnormal findings. Demographic histogram of 32 CTPA abnormalities extracted from reports using LLM.

using either organ-specific or abnormality-specific prompting, followed by report composition assigned with clinical reporting standards. The performance of various CTPA reading agents under these strategies is summarized in Fig. 4. To assess report quality, we applied standard natural language generation (NLG) metrics including BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr, and BERT-F1. These metrics evaluate lexical overlap, fluency, and semantic similarity between generated reports and ground-truth references.

In the caption-based setting, CT-CHAT achieved a BERT-F1 of 0.856 (95% confidence intervals [CI]: 0.855–0.856) and BLEU-4 of 0.142 (95% CI: 0.140–0.144) on the BUH dataset under the "Caption + Organ List + One-shot" setting. This compared to 0.830 and 0.069 for RadFM, and 0.770 and 0.001 for M3D, respectively ($p < 0.001$).

In the region-based generation setting, strategies guided by predicted abnormalities showed higher metric scores than those using organ-only prompts. On the BUH dataset, the abnormality-predicted (Abn-Pred) strategy using CT-CHAT achieved a BERT-F1 of 0.874 (95% CI: 0.873–0.874) and BLEU-4 of 0.149 (95% CI: 0.146–0.152). The Abn-Pred outperforms the abnormality-all (Abn-ALL) variant without abnormality-informed, which yielded BERT-F1 of 0.832 and BLEU-4 of 0.037. The upper-bound condition using ground-truth abnormality labels (Abn-GT) produced the highest scores, which reached BERT-F1 of 0.881 and BLEU-4 of 0.169.

Table 1 | Patient characteristics of the total and prognostic cohorts

	Brown University Health System	Johns Hopkins University-affiliated hospital	*p* value
CTPA-report paired data			
Patient number	19565	1077	
CTPA image number	59754	5759	
Sex			0.6143
Male	8327 (42.6%)	464 (43.1%)	
Female	11238 (57.4%)	613 (56.9%)	
Age (years)	60.0 (18.0)	57.0 (26.0)	<0.0001
Multimodal prognosis data			
PE patient number	917	95	
Sex			0.9142
Male	435 (47.4%)	46 (48.4%)	
Female	482 (52.6%)	49 (51.6%)	
Age (years)	64.0 (25.0)	58.0 (27.5)	0.1063
PESI variables			
Age ≥ 80	177 (19.3%)	14 (14.7%)	0.3353
Chronic cancer	262 (28.6%)	43 (45.3%)	0.0014
Chronic heart failure	70 (7.6%)	27 (28.4%)	<0.0001
Chronic obstructive pulmonary disease	195 (21.3%)	28 (29.5%)	0.0696
Heart rate ≥ 110 beats/min	149 (16.2%)	8 (8.4%)	0.0518
Systolic BP < 100 mmHg	86 (9.4%)	12 (12.6%)	0.2793
Respiratory rate ≥ 30 breaths/min	24 (2.6%)	48 (50.5%)	<0.0001
Temperature < 96. 8 °F	41 (4.5%)	17 (17.9%)	<0.0001
Altered mental status	66 (7.2%)	25 (26.3%)	0<0.0001
O2 saturation < 90%	24 (2.6%)	12 (12.6%)	<0.0001
PESI score	87.0 (44.0)	122 (67.5)	<0.0001
Death	163 (17.8%)	46 (48.4%)	<0.0001
Follow up days	1212.0 (1594.0)	404.5 (559.0)	<0.0001
Short-term follow up (days < 30)	124 (13.5%)	7 (7.4%)	0.1086

The total cohort comprises CTPA image-report pairs collected from two academic institutions. Demographic information and PESI variables are summarized for the prognostic subset. Continuous variables are presented as median (interquartile range), and categorical variables as counts (percentage). *P* values are calculated based on comparisons between the two cohorts.

Compared to the caption-based CT-CHAT baseline for holistic generation, the Abn-Pred strategy for structured generation improved BERT-F1 by 2.1% (from 0.856 to 0.874) and BLEU-4 by 4.9% (from 0.142 to 0.149). In addition, ROUGE-L, as a metric reflecting content recall and structural alignment, also increased by 16.3% (from 0.245 [95% CI: 0.243–0.246] to 0.285 [95% CI: 0.282–0.288]).

Across datasets, structured report generation using the BUH-designed framework showed variations in performance due to differences in institutional report styles. For instance, BLEU-4 scores were lower on INSPECT and JHU datasets, while other metrics such as BERT-F1 and ROUGE-L showed moderate changes. The INSPECT dataset includes only brief, unstructured "Impression" sections, which limits direct comparison with full structured reports and affects alignment on language-based metrics.

CT-CHAT under the "Caption + Organ List + One-shot" setting yielded the highest scores among caption-based baselines. Our structured generation approaches, particularly those incorporating abnormality predictions, reported higher metric scores across all three datasets.

Study findings comparison

To generate study-level radiology reports, a report-writing agent was used to aggregate region-level findings into structured "Study Findings" sections. As shown in Fig. 5a, we compared the performance of structured generation methods, specifically organ-based and abnormality-based approaches, with the strongest caption-based baseline ("Caption + Organ List + One-shot"). This baseline leverages organ-specific cues and example-driven prompting to improve long-form reasoning and contextual alignment.

Across all testing datasets and CTPA reading agents (CT-CHAT, RadFM, M3D), the abnormality-based strategy (Abn-Pred) yielded higher scores than the caption-based baseline. On the BUH dataset using CT-CHAT agent, the Abn-Pred strategy achieved BERT-F1 of 0.880 (95% CI: 0.879–0.880], which improved by 1.3% over captioning, and BLEU-4 by 114.9% (from 0.074 to 0.159). ROUGE-L also increased by 43.8%, and METEOR by 28.8% (from 0.160 to 0.206). CIDEr score showed a relative improvement of 26.3% (from 0.019 to 0.024).

The abnormality-guided generation demonstrated superior generalizability in cross-institutional settings. On the INSPECT dataset, Abn-Pred using the CT-CHAT agent achieved a BERT-F1 of 0.829 (95% CI: 0.829–0.830], outperforming the captioning approach (0.828). This performance gap ($p < 0.001$) widened with M3D (0.829 vs. 0.825) and RadFM (0.830 vs. 0.824). On the JHU dataset, Abn-Pred show stronger language-grounding with BERT-F1 values of 0.842 (CT-CHAT), 0.843 (M3D), and 0.843 (RadFM). Despite lower absolute scores in BLEU-4 due to the unstructured nature of the INSPECT reports (BLEU-4 = 0.001 for Abn-Pred), BERT-F1 remained the highest (0.828-0.830) across all agents for Abn-Pred.

Compared to organ-based generation, the Abn-Pred strategy consistently outperformed across all three agents on the BUH dataset. For CT-CHAT, BERT-F1 increased 3.8% (from 0.858 to 0.891), 1.9% for RadFM, and 4.2% for M3D, respectively. In addition to BERT-F1, Abn-Pred also led to consistent gains in BLEU-4, ROUGE-L, and CIDEr scores compared to organ-guided prompts.

While using ground-truth abnormality labels (Abn-GT), the upper-bound performance was observed across datasets. On BUH, Abn-GT achieved the highest values across all metrics, including BERT-F1 of 0.897, BLEU-4 of 0.179, and ROUGE-L of 0.337.

Study impression comparison

To generate the "Study Impression" section, the report-writing agent synthesized high-level diagnostic summaries from the structured "Study Findings", focusing on key conclusions such as the presence or absence of PE. As shown in Fig. 5b, we evaluated the same four prompting strategies used in the findings generation task: holistic captioning, organ-based, and two abnormality-based methods (Abn-Pred and Abn-GT).

Across datasets and model configurations, abnormality-based prompting consistently outperformed holistic caption-based approaches. On the BUH dataset, the Abn-Pred prompting strategy consistently outperformed the caption-based baseline across all agents. For CT-CHAT, the BLEU-4 score tripled from 0.005 to 0.015 (95% CI: 0.015–0.016), while ROUGE-L increased by over 5% from 0.172 to 0.181. BERT-F1 remained stable at 0.879–0.880 across both settings. Notably, RadFM and M3D also exhibited substantial gains in BLEU-4, increasing from 0.005 to 0.018 and 0.016, respectively.

On the INSPECT dataset, where heterogeneous reporting styles contributed to lower absolute scores, the Abn-Pred strategy consistently outperformed the caption-based baseline across all agents. M3D agent achieved BLEU-4 of 0.011 (95% CI: 0.010–0.011), improved from 0.003, and ROUGE-L increased by 12.9% from 0.139 to 0.157, while BERT-F1 remained comparable (0.825 vs. 0.822). All improvements were statistically significant ($p < 0.001$). Similar improvements were observed over 50% improved BLEU-4 for CT-CHAT and RadFM.

In the JHU cohort, the Abn-Pred strategy using CT-CHAT achieved a BLEU4 score of 0.016 (95% CI: 0.015–0.018), representing a 167% increase over the caption-based baseline (0.006; 95% CI: 0.005–0.007). ROUGE-L remained comparable (0.152 vs. 0.152, $p = 0.002$), and BERT-F1 showed a slight decrease from 0.849 to 0.838 ($p < 0.001$). Consistent trends were

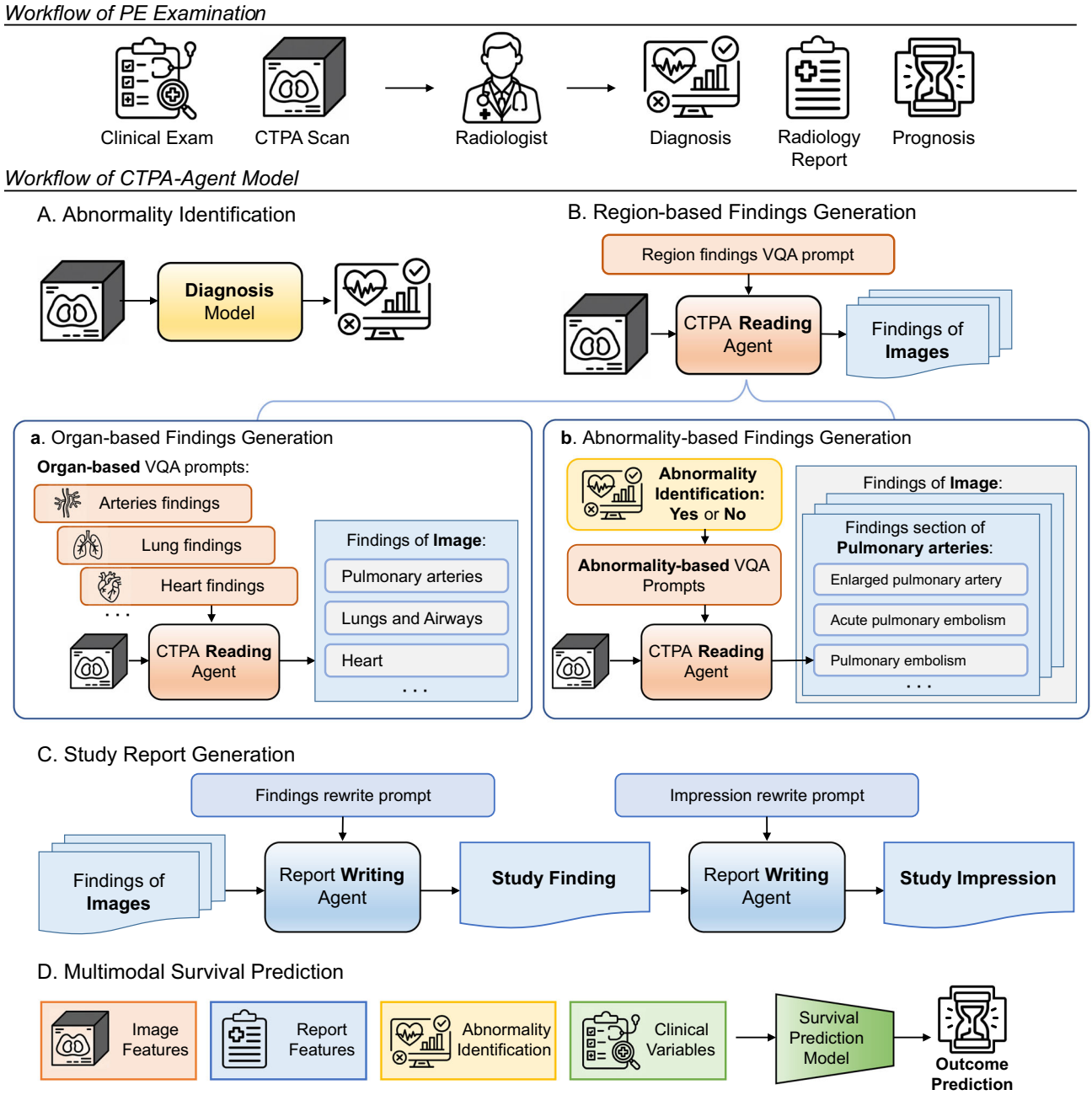


Fig. 2 | Overview of agent-based CTPA examination approach. The workflow between radiologists and the proposed CTPA examination pipeline includes disease diagnosis, report generation, and prognosis assessment. **A** The diagnosis module identifies 32 PE-related abnormalities. **B** The CTPA Reading Agent processes vision-driven queries to extract region-specific findings and generate image-level reports, prompted to interpret organ-based findings (**a**) and abnormality-based findings (**b**). **C** The Report Writing Agent synthesizes these findings into a diagnostic report. **D** The multimodal survival prediction module estimates patient survival risk.

observed with other agents, RadFM and M3D. The Abn-GT, which uses ground-truth abnormality labels, yielded the highest metric values across datasets, including a BERT-F1 of 0.887 on BUH and BLEU-4 of 0.017 on JHU.

For organ-based prompting, scores were also high but slightly lower than those obtained using abnormality-based methods. For example, on BUH with RadFM, the organ-based method achieved a BERT-F1 of 0.878 and CIDEr of 0.020, while Abn-Pred achieved 0.882 and 0.011, respectively ($p < 0.001$).

Case study and visualization: organ-based vs. abnormality-informed generation

To illustrate the outputs of our organ-based and abnormality-informed report generation strategies, we selected two representative CTPA cases and compared the generated reports with the corresponding ground truth. As shown in Fig. 6, the structured findings are visualized, with blue text indicating agreement with ground truth and red text denoting inaccurate or irrelevant content.

In the first case, the abnormality-informed method generated findings that included "no pulmonary embolism," "bilateral pleural effusions," "bilateral lower lobe volume loss", and "normal mediastinum and hila," all of which matched the ground-truth report. In contrast, the organ-based method produced additional findings that are not present in the reference, including minor or unrelated observations.

In the second case, the abnormality-informed method correctly included findings such as "no pulmonary embolism", and "diffuse thickening of the bronchial wall". The generated phrase "calcification of the aortic

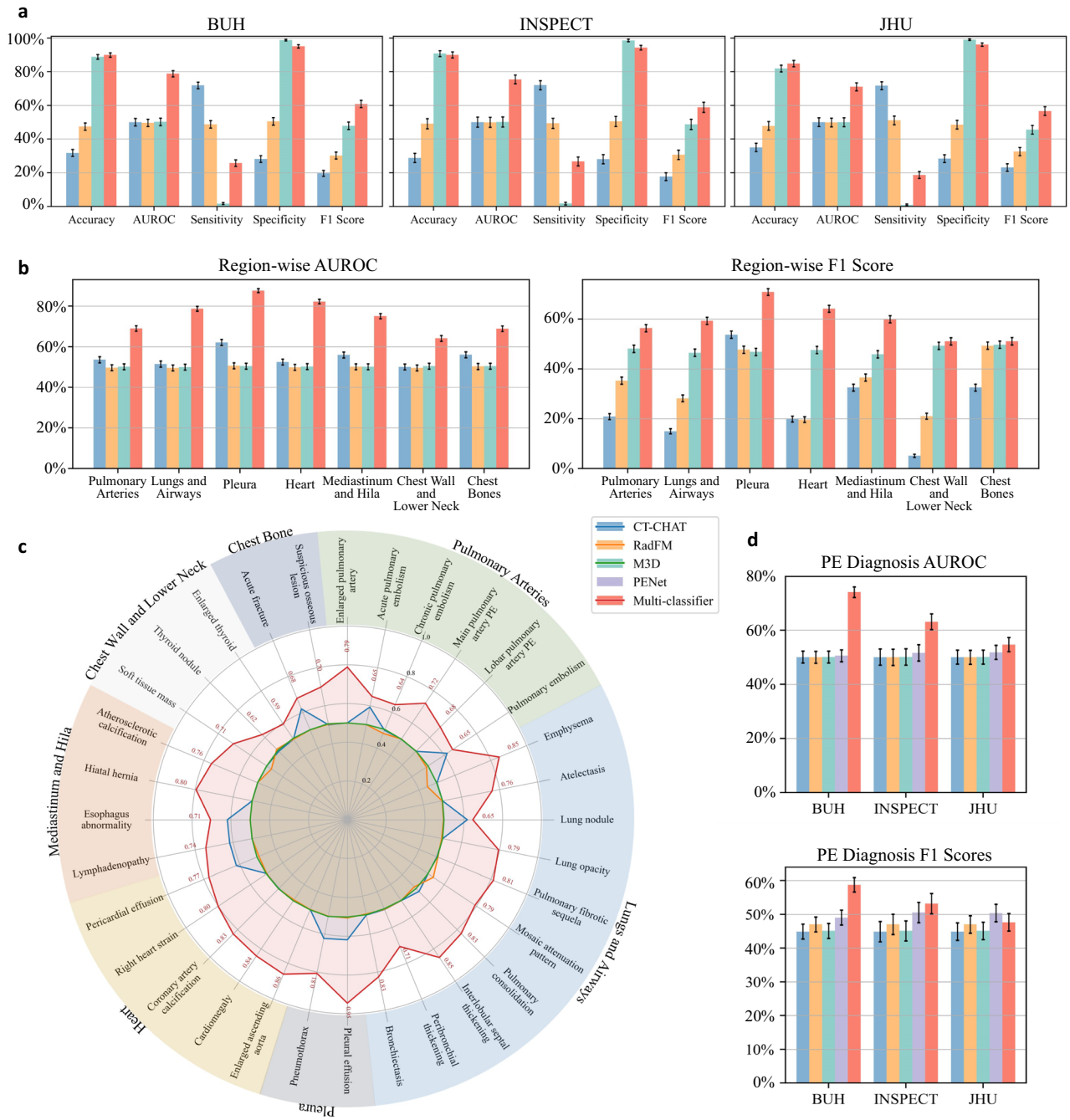


Fig. 3 | Comparison of abnormality identification performance. a Multi-abnormality classification performance on three datasets (BUH, INSPECT, JHU) evaluated by Accuracy, AUROC, Sensitivity, Specificity, and F1 Score. **b** Region-wise evaluation of AUROC and F1 Score for abnormalities across seven anatomical regions (e.g., Pulmonary Arteries, Lungs and Airways, Pleura). **c** Per-abnormality AUROC performance on 32 types of CTPA related findings, organized by anatomical regions. **d** Performance comparison of PE diagnosis evaluated by AUROC and F1 Score across the three datasets, including results from the PE-specific detection network PENet.

wall" corresponded to the reference phrase "scattered atherosclerotic calcification," although not an exact match. The organ-based method, in this instance, included findings not supported by the reference report, such as "aortic dissection with an intimal flap" and "well-defined nodule with a halo sign".

Expert evaluation: compared with holistic captioning

To assess the clinical quality of generated radiology reports beyond automatic metrics, we conducted a blinded expert evaluation involving three independent review groups led by board-certified radiologists. The evaluation compared two generation strategies: (1) a holistic caption-based method ("Caption + Organ List + One-shot"), and (2) our structured, abnormality-guided method (Abn-Pred). Both methods used CT-CHAT as the CTPA reading agent and Llama 3 as the report-writing agent.

From the BUH testing set, 30 patient cases were randomly selected. For each case, reviewers were presented with the ground-truth report alongside two anonymized generated versions—one from each method—for both the "Study Findings" and "Study Impression" sections. Reviewers were blinded to the generation strategy and asked to choose the version with higher clinical quality based on accuracy, clarity, and relevance to the reference report. Each evaluation included a 5-point confidence score (1 = lowest, 5 = highest) to indicate the strength of preference.

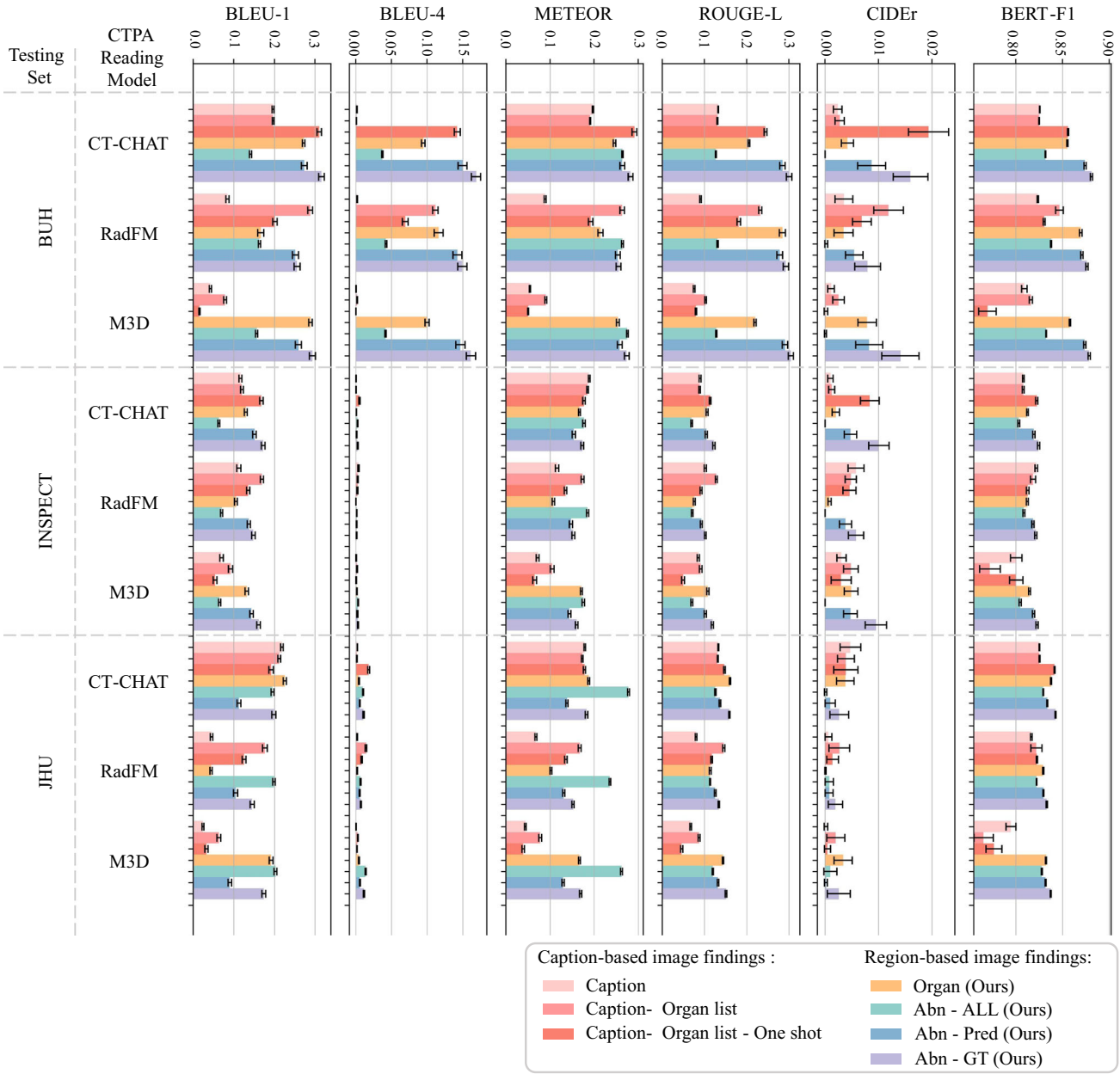


Fig. 4 | Comparison of image-level report generation performance. We compare two strategies for generating radiology reports: (1) Caption-based generation directly produces holistic reports, including three variants: caption-only, caption with organ list, and one-shot caption with organ list. (2) Region-based generation first extracts region-wise findings (either organ or abnormalities) and then assembles them into structured reports based on the hierarchical reporting framework. Region-based variants include organ-wise findings, abnormality-wise findings (Abn-All), predicted abnormalities informed (Abn-Pred), and ground-truth abnormalities informed (Abn-GT).

As shown in Fig. 7, the structured generation method was preferred in the majority of cases across all evaluators. For the "Study Findings" section, Abn-Pred was selected in 90.0%, 80.0%, and 96.7% of cases by Experts 1, 2, and 3, respectively. For the "Study Impressions" section, it was selected in 90.0%, 83.3%, and 93.3% of cases. Report selections were consistently associated with higher confidence scores.

Multimodal prognosis performance

We evaluated the performance of the multimodal survival prediction module using Concordance index (C-index) as a measure of agreement between predicted risk scores and observed outcomes[25]. Higher C-index values indicate greater predictive accuracy, with values closer to 1 representing near-perfect concordance. As shown in Fig. 8a, integration of image-derived and clinical information improved prediction performance compared to traditional or single-modality approaches.

In the BUH cohort, the PESI baseline achieved a C-index of 0.764. Among single-modality models, clinical variables (Clin) performed best with a C-index of 0.789, followed by diagnosis-based features (Dia: 0.781), text from generated radiology reports (Text: 0.786), and imaging features (Img: 0.751). Multimodal fusion yielded higher C-index values: Img + Text (0.798), Img + Clin (0.817), and Img + Clin + Text (0.846). The highest performance was achieved with the full combination of Img + Clin + Dia + Text, reaching a C-index of 0.863.

In the JHU cohort, lower overall performance was observed. PESI achieved a C-index of 0.596, while Img (0.635), Text (0.606), and Clin (0.571) showed modest predictive value. Multimodal fusion improved performance: Img + Clin + Text achieved a C-index of 0.719, and the best

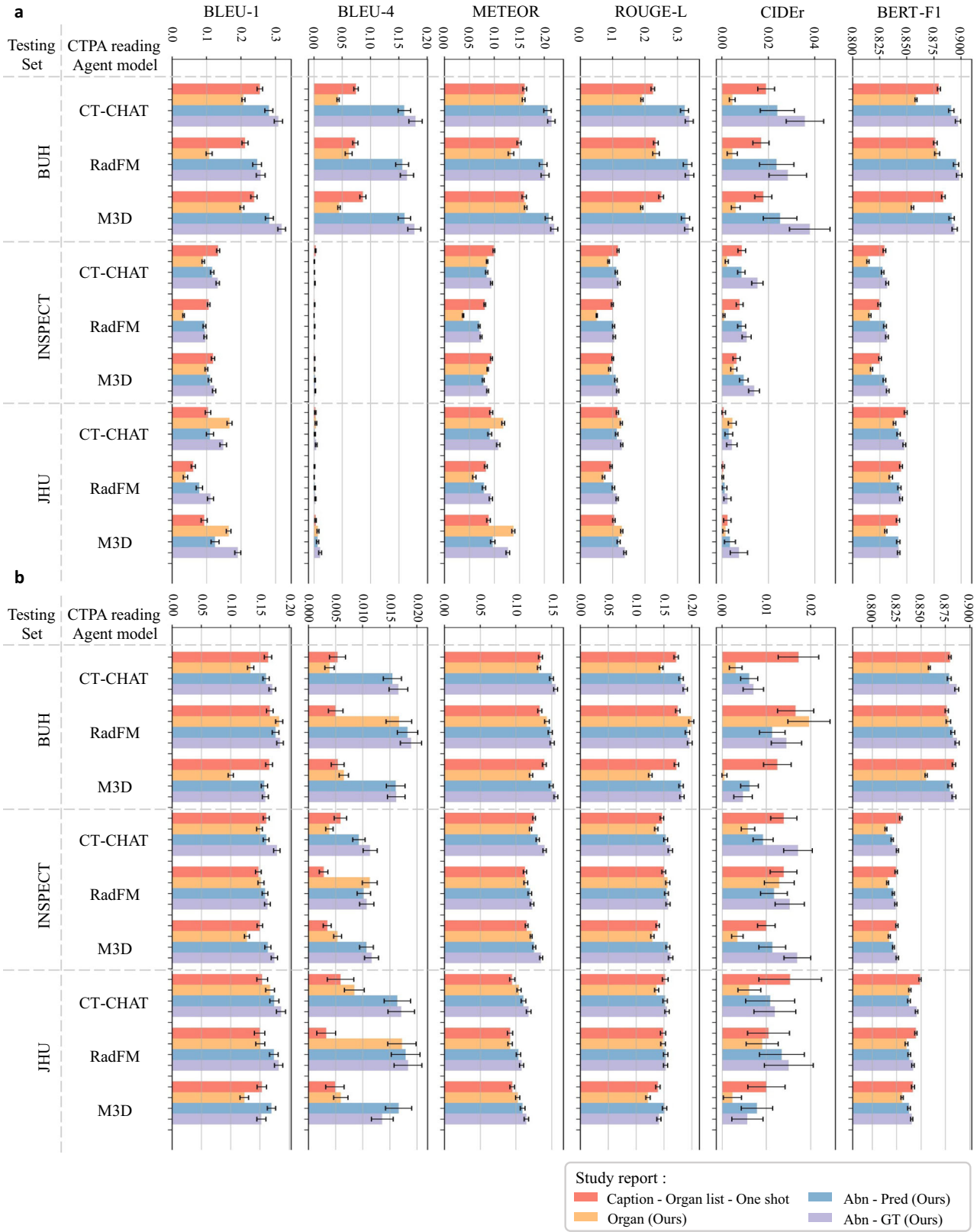
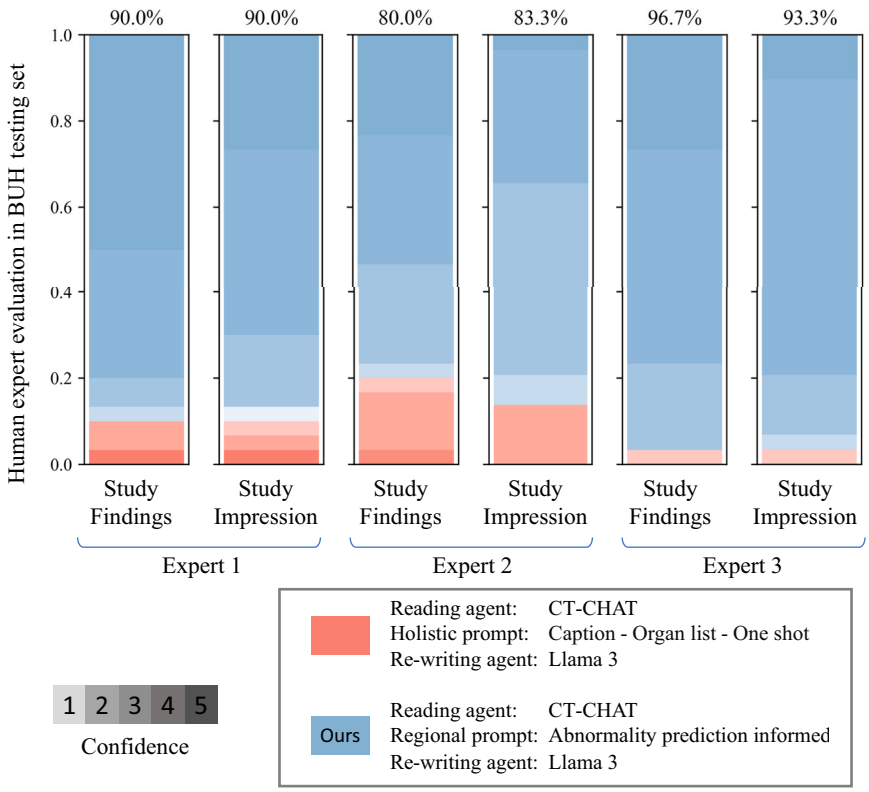


Fig. 5 | Comparison of study-level report generation performance. The LLM with prompting strategies aggregates and summarizes image-level findings across multiple scans within a single study to generate comprehensive diagnostic sections. **a** Performance of generating the Findings section, which consolidates abnormality-level descriptions from individual image-level reports. **b** Performance of generating the Impression section, which synthesizes the key clinical conclusions derived from the aggregated study Findings.

Ground Truth	Organ-based (Ours)	Abn-Pred (Ours)
Study 1: FINDINGS: Pulmonary arteries: *No pulmonary emboli are identified.* Lungs and Airways: *Bilateral lower lobe volume loss* Pleura: *Bilateral pleural effusions* considered small. Heart: *Normal.* Mediastinum and Hila: *Normal.* Chest Wall and Lower Neck: *Normal.* Chest Bones: *Normal.*	FINDINGS: Pulmonary arteries: Large saddle embolus is present. Lungs and Airways: Extensive pulmonary parenchymal and interstitial fibrosis with reticular opacities, associated with marked dilatation of pulmonary veins and ascites. Pleura: *Pleural effusion* is present on the left side. Heart: Aortic dissection with a large false lumen on the right, compressing the smaller true lumen, and moderate aortic valve calcification. Mediastinum and Hila: Enlarged lymph nodes are present. Chest Wall and Lower Neck: A simple, soft tissue density aneurysm is present in the proximal portion of the left subclavian artery. Chest Bones: *No acute abnormality is present.*	FINDINGS: Pulmonary arteries: *Normal.* Lungs and Airways: *Bilateral atelectasis.* Pleura: *Bilateral pleural effusion*, with significant thickening of the parietal pleura on the left side, and a unilateral pneumothorax. Heart: Moderate-sized pericardial effusion with thickening. Mediastinum and Hila: *Normal.* Chest Wall and Lower Neck: *Normal.* Chest Bones: *Normal.*
Study 2: FINDINGS: Pulmonary arteries: *No pulmonary emboli are identified.* Lungs and Airways: *Diffuse bronchial wall thickening*. Background of *emphysematous* changes. Pleura: *Normal.* Heart: *Normal.* Mediastinum and Hila: *Scattered atherosclerotic calcification* seen in aorta and branching vessels. No bulky lymphadenopathy. Chest Wall and Lower Neck: *Normal.* Bones: *No acute abnormality.*	FINDINGS: Pulmonary arteries: Chronic pulmonary embolism. Lungs and Airways: No significant abnormality is identified in the lungs and airways. Pleura: Pleural effusion on the left side. Heart: Cardiomegaly. Mediastinum and Hila: Normal thymus tissue is present in the anterior mediastinum. Chest Wall and Lower Neck: Diffuse irregular thickening of the chest wall and lower neck with regions of necrosis. Chest Bones: Bilateral first rib cortical thinning/sclerosis with posterior kyphosis.	FINDINGS: Pulmonary arteries: *Normal.* Lungs and Airways: Well-defined nodule with a halo sign, *diffuse thickening of the bronchial walls*. Pleura: *Normal.* Heart: Aortic dissection with intimal flap. Mediastinum and Hila: *Calcification of the aortic wall.* Chest Wall and Lower Neck: *Normal.* Chest Bones: *Normal.*

Fig. 6 | Examples of the generated reports. The proposed organ-based and abnormality-based results are compared with the structured ground truth. The blue italic text is the correct predictions corresponding to the actual reports, and the red areas indicate the untrue information in the predictions.

Fig. 7 | Human expert evaluation of generated study reports on the BUH testing set. Three independent reviewing groups led by board-certified radiologists independently evaluated the quality of generated Study Findings and Study Impression sections, comparing outputs from two prompting strategies: a holistic caption-based method ("Caption + Organ list + One-shot") and our proposed structured generation approach informed by abnormality predictions. All generations were produced using CT-CHAT as the reading agent and LLaMA 3 as the report-writing agent. For each report pair, radiologists selected the version with higher clinical quality, referencing the ground truth report as context. Stacked bars represent the normalized distribution of preference scores across five levels of confidence (1 = least confident, 5 = most confident).



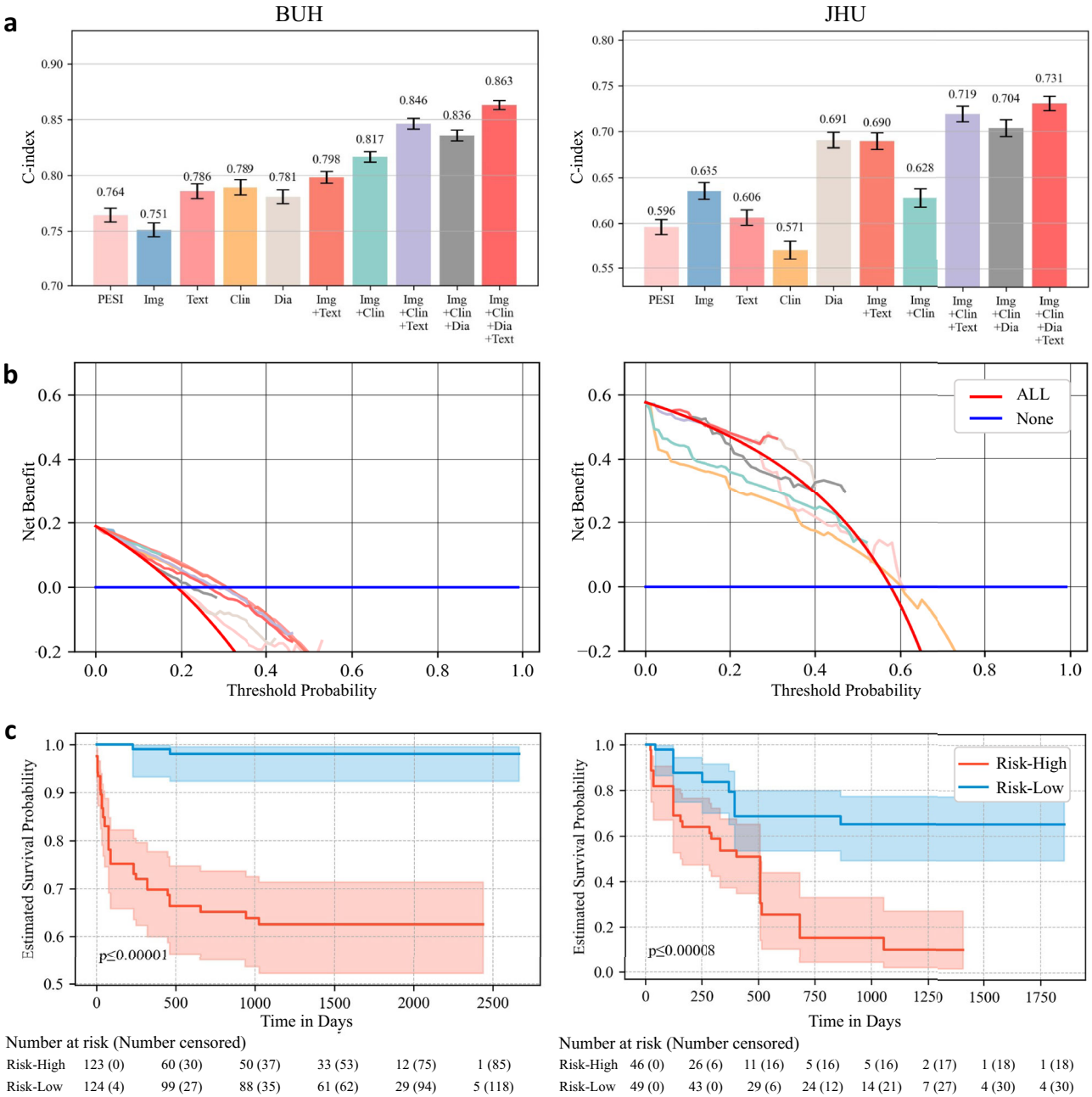


Fig. 8 | Performance comparison of multimodal survival prediction modules. **a** Concordance index (C-index) for different combinations of modality inputs, including PESI scores, imaging (Img), clinical variables (Clin), diagnosis (Dia), and generated report text (Text), across the BUH and JHU testing cohorts. Multimodal fusion models consistently outperform unimodal baselines. **b** Decision Curve Analysis (DCA) of multimodal survival prediction modules illustrates the net clinical benefit of various unimodal and multimodal models across a range of threshold probabilities for risk stratification. **c** Kaplan-Meier survival curves for high-risk and low-risk groups stratified by the median risk score from the four-modal fusion model (Img + Clin + Dia + Text).

performance again came from the full four-modal model (C-index: 0.731), confirming the generalizability of our fusion strategy across datasets.

To evaluate potential clinical utility, we performed Decision Curve Analysis (DCA)[26], shown in Fig. 8b. DCA evaluates the net benefit of different predictive models across a range of threshold probabilities, reflecting their potential impact on clinical decision-making. In both cohorts, the full four-modal model (Img + Clin + Dia + Text) consistently yielded the highest net benefit across a broad range of thresholds, surpassing unimodal and partial multimodal baselines. This advantage was particularly pronounced in the JHU cohort with lower censoring rate, where traditional clinical risk scores (e.g., PESI) and single-modality predictors demonstrated limited utility.

Kaplan-Meier survival analysis[27] based on the predicted risk scores from the multimodal model are shown in Fig. 8c. These curves demonstrate significant differences in survival probabilities over time between high-risk and low-risk groups in both cohorts (log-rank test: $p < 0.00001$ for BUH, $p < 0.00008$ for JHU), underscoring the model's effectiveness in patient risk stratification.

Discussion

In this study, we introduced a novel framework for CTPA image report generation, leveraging a large-scale, retrospective dataset. The results demonstrate that incorporating abnormality detection into the reporting pipeline enables strong performance across multiple clinical tasks, including

structured radiology reporting and multimodal survival prediction. These findings also highlight key limitations, providing direction for future improvements in automated radiology reporting systems.

Despite the potential of medical VLMs in 3D imaging, current approaches struggle to produce fully comprehensive, structured reports that synthesize findings across multiple anatomical regions. Our experiments showed that these models frequently miss or inadequately describe a wide range of abnormalities, leading to incomplete or insufficiently detailed reports that limit clinical utility. Furthermore, medical VLMs demonstrate limited ability to construct coherent and generalized narratives from complex prompts, particularly when summarizing diagnostic impressions. Their lower responsiveness compared to LLMs highlights the need for new strategies that more effectively link diagnostic reasoning with generative capabilities.

To address these limitations, we developed a region-based generation approach embedded within a structured, clinically grounded diagnostic framework tailored for CTPA interpretation. The framework mimics radiologists' diagnostic workflow by organizing analysis around seven anatomically defined regions and 32 clinically significant abnormalities. This hierarchical structure supports improved diagnostic accuracy, enhances report clarity, and aligns with established clinical guidelines. At the core of the framework is a multi-label abnormality classification module that detects findings across anatomical regions, serving both as a diagnostic endpoint and as input for region-level report generation. A CTPA reading agent utilizes abnormality-specific prompts to extract targeted diagnostic content, which is then synthesized by a report-writing agent into structured output. This includes a "Study Findings" section summarizing region-wise observations and an "Impression" section highlighting key diagnostic conclusions. Additionally, a multimodal survival prediction module integrates imaging, clinical variables, diagnostic outputs, and generated reports to estimate patient risk, thereby extending the utility of the framework to support outcome prediction and pulmonary embolism (PE) management.

We evaluated the proposed framework across three core tasks: abnormality classification, radiology report generation, and survival prediction. Each task was benchmarked against SOTA medical VLMs for 3D imaging, including CT-CHAT[11], RadFM[12], and M3D[13]. Evaluation protocols were aligned with the specific objectives of each task to ensure fair and relevant comparisons.

For abnormality classification, our multi-label classifier achieved higher accuracy in detecting 32 co-occurring CTPA abnormalities compared to prompt-based visual question answering strategies used in VLMs. This performance advantage stems from the classifier's ability to capture region-specific visual features and model label co-occurrence, which enhances both sensitivity and precision across diverse anatomical regions.

For the report generation task, we compared organ-based and abnormality-based structured prompting methods against strong holistic captioning baselines (e.g., "Caption + Organ List + One-shot"). Across the BUH, INSPECT, and JHU datasets, abnormality-guided prompting consistently outperformed both holistic and organ-based methods on BLEU-4, ROUGE-L, METEOR, BERT-F1, and CIDEr metrics. These results indicate that incorporating diagnosis-aware input enhances the generation of semantically relevant and clinically aligned content. Abnormality-guided reports were also more concise and better conformed to radiology reporting standards.

Expert evaluation supported these findings. Three board-certified radiologists compared reports generated by structured prompting and holistic caption-based methods, consistently preferring the abnormality-guided reports. They cited better diagnostic alignment, clearer articulation of PE-related findings, and higher clinical relevance in both findings and impression sections. In addition to higher lexical and semantic scores, radiologists emphasized several qualitative advantages of the abnormality-based framework. Specifically, these reports demonstrated improved clinical clarity, reduced redundancy, and sharper diagnostic focus. The structured format enabled a more systematic presentation of abnormal findings, which better mirrored radiologists' reasoning processes during image interpretation. The report-writing agent behind these outputs prioritized clinically significant content and emphasized PE-related conclusions, leading to more actionable impressions and a closer match to expert-level reporting expectations. These qualitative insights were consistently cited during evaluation and underscore the utility of abnormality-guided prompting beyond mere performance metrics.

For survival prediction, our four-modal fusion model, which integrates imaging, predicted abnormalities, structured reports, and clinical variables, achieved the highest C-index in both BUH and JHU cohorts, outperforming unimodal models and the conventional PESI score. Decision curve analysis confirmed the model's clinical utility, demonstrating consistently higher net benefits across a range of decision thresholds. These findings highlight the advantage of incorporating structured textual outputs alongside multimodal data for personalized risk stratification.

Our framework builds upon recent advancements such as CT-RATE[11] and INSPECT[14], integrating the strengths into a unified, multi-agent system. This design addresses the limitations of single-model VLMs by introducing a structured, diagnosis-aware generation strategy that reduces redundancy, emphasizes clinically significant findings, and enables robust survival prediction. The generated reports are more consistent with clinical reporting standards, supporting improved diagnostic accuracy and downstream decision-making. Additionally, the framework provides a scalable solution that minimizes false positives and missed abnormalities, reduces reporting variability, and facilitates clinically meaningful risk stratification for patients with pulmonary embolism.

Despite its strengths, this study has several important limitations that warrant consideration. First, the structured reporting framework relies on a predefined set of 32 PE-related abnormalities. While this promotes diagnostic standardization, it limits the system's capacity to identify incidental or atypical findings-such as cardiovascular anomalies, extracardiac pathology, or subtle parenchymal changes-that may carry clinical significance. Second, the retrospective study design, though leveraging large-scale multimodal data, introduces potential biases related to data quality, documentation variability, and institutional heterogeneity. These factors may affect model generalizability to prospective use, alternative imaging protocols, and underrepresented populations or rare conditions.

Nevertheless, the framework is inherently extensible. The abnormality hierarchy can be customized to reflect different institutional standards or expanded to cover broader disease domains. Additionally, the classifier can be retrained with new labeled data, supporting scalable adaptation and deployment. Although our primary focus was evaluating region-level report generation using a state-of-the-art I3D backbone from Merlin[28], we acknowledge that ablation studies on backbone architectures and pretraining strategies are needed to optimize performance. Exploring alternative visual encoders and domain-specific pretraining remains an important direction for future work.

In the survival prediction task, the low event rate in the external JHU cohort may have reduced the statistical power of the Cox regression model, potentially limiting the reliability of risk estimates. This challenge is especially pertinent to real-world deployment, where accurately predicting rare but clinically critical outcomes is essential. Furthermore, while combining VLMs and LLMs improves report coherence and interpretability, current LLMs still struggle to synthesize nuanced hierarchical content and may overlook clinically prioritized findings.

Limitations also exist in our evaluation methodology. Traditional natural language generation metrics such as BLEU and ROUGE emphasize lexical overlap rather than clinical accuracy, potentially underestimating the true utility of generated reports. Emerging domain-specific metrics like RadGraph[29] offer more diagnostic relevance but require further validation and broader adoption. Expert evaluation, while clinically informative, also has constraints. Our study lacks a direct comparison against board-certified radiologists in abnormality detection and survival prediction tasks, making it difficult to gauge whether the AI system achieves expert-level performance. Additionally, while radiologists preferred abnormality-based reports over organ-based ones, our analysis did not fully elucidate the rationale behind this preference. A more comprehensive framework that

incorporates both quantitative benchmarking and structured qualitative feedback would enhance understanding of clinical impact.

To support clinical translation, future work should include prospective validation studies to assess real-time diagnostic and prognostic performance, interpretability, and safety of the framework. In particular, benchmarking against board-certified radiologists in real-world workflows-especially in cases with incidental findings or atypical presentations, is crucial to establish clinical credibility and adoption.

Methods

In this section, we present the proposed multi-class abnormality classification module and the abnormality-guided, region-based report generation framework, developed to enable comprehensive evaluation of pulmonary embolism (PE) across key clinical tasks. The model was validated on two internal datasets and one publicly available dataset, ensuring diverse and representative cohort coverage for robust performance assessment.

Patient cohorts

This retrospective study included patients who underwent CTPA scans at two major academic medical centers: Brown University Health (BUH) and Johns Hopkins University (JHU). The BUH dataset comprises 59,754 paired CTPA image-report studies from 19,565 patients who received scans between 2015 and 2019 at Rhode Island Hospital, The Miriam Hospital, and Newport Hospital. This dataset was used for both abnormality identification and radiology report generation tasks, with data split into training, validation, and testing sets in a 7:1:2 ratio. For prognosis evaluation, we selected a subset of 917 patients with confirmed PE from a previously published study[30], including corresponding CTPA images, radiology reports, 11 PESI variables derived from electronic health record (EHR), and outcome data (mortality status and follow-up duration days).

For external validation on three clinical tasks, we curated a dataset from Johns Hopkins Hospital comprising 5759 CTPA scans from 1077 patients collected between 2016 and 2020. Among these, 95 patients with confirmed PE and complete outcome records were included for multimodal survival analysis using the same inclusion criteria as the BUH cohort.

The study protocols were reviewed and approved by the Lifespan Institutional Review Board 3 (covering Rhode Island Hospital, The Miriam Hospital, Newport Hospital, and affiliated institutions; reference number: 1791856-20, project code: 214421) and the Johns Hopkins Medicine Institutional Review Board (reference number IRB00424745). The requirement for informed consent was waived by both ethics committees, as the study involved retrospective analysis of de-identified imaging and clinical data that were either publicly available or recorded in a manner that precluded identification of individual subjects. All patients included in the analysis were over 18 years of age.

The INSPECT dataset[14] serves as the largest publicly available CTPA dataset to date. It contains 23,248 CTPA scans from 19,402 patients collected at Stanford Medicine between 2000 and 2021. Notably, it includes only the "Impression" sections of radiology reports-brief, unstructured summaries written by expert radiologists-rather than full structured reports. This may inherently limit the evaluation of report completeness and detailed findings. We randomly sampled one-fifth of the data to construct an external validation dataset for abnormality diagnosis and report generation validation. Figure 1a illustrates the study design and patient inclusion criteria. Table 1 summarizes the demographic and clinical characteristics across BUH and JHU datasets. Data collection and analysis were performed locally at each center, with strict measures to maintain patient anonymity.

Multi-class abnormality diagnosis

We propose a multi-label classification module to detect 32 clinically significant abnormalities in CTPA scans, employing a 3D inflated ResNet-152 (I3D) as the backbone. The 2D convolutional structureare of ResNet-152 inflated into 3D, enabling effective volumetric feature extraction through 3D convolutional kernels tailored for 3D medical imaging. To enhance feature representation, the model is initialized with pretrained weights from Merlin[28], a vision-language foundation model tailored for 3D CT analysis and pretrained on structured EHR data and unstructured radiology reports. This initialization enables the model to capture spatially and semantically enriched pathological features. The network begins with a 7×7 in-plane convolution followed by a 3×3 max-pooling layer, reducing the input resolution by a factor of four. Subsequent features are aggregated via average pooling and a $1 \times 1 \times 1$ Conv3D layer to produce probability estimates for the 32 PE-related abnormalities. The diagnosis module supports multi-label predictions per region and distinguishes between co-occurring and visually subtle abnormalities.

We compared our classifier against leading 3D medical VLMs: CT-CHAT[11], RadFM[12], and M3D[13]. These models operate under a visual question answering (VQA) paradigm. Each VLM was queried per abnormality using the prompt:

"Is there any indication of < *Abnormality* > in this image? (This is a true or false question, please answer 'Yes' or 'No')."

This setup enabled direct comparison between structured classification and prompt-based querying.

Region-based report generation

To generate structured radiology reports, we implemented a region-based report generation framework guided by anatomical and abnormality prompts. As illustrated in Fig. 2B, the pipeline includes a CTPA reading agent, powered by a 3D medical VLM, which is prompted to generate region-wise findings based on two prompting strategies:

- The organ-based finding captioning strategy follows clinical reporting conventions by prompting findings by organ region, as illustrat Fig. 2a. For each of the seven predefined anatomical regions including pulmonary arteries, lungs and airways, pleura, heart, mediastinum and hila, chest wall and lower neck, and chest bones, the CTPA reading agent was queried with:

"What findings of < *Organ* > do you observe in this medical image?"

This method captures both normal and abnormal findings and maintains spatial coherence. However, without explicit diagnostic cues, it can produce generic descriptions or miss subtle but important abnormalities.

- The abnormality-based findings captioning strategy enhances diagnostic precision by leveraging 32 predefined abnormalities to guide the prompting process, as illustrated in Fig. 2b. This method leverages the outputs from the abnormality classifier to guide CTPA reading agent with targeted VQA queries:

"What findings of < *Abnormality* > do you observe in this medical image?"

Only abnormalities predicted as present are queried, focusing generation on clinically relevant content while suppressing irrelevant responses. This strategy improves specificity and enables the model to describe imaging features such as acute PE, lymphadenopathy, or parenchymal changes in a concise and interpretable format.

Structured regional outputs from either the organ-based or abnormality-based methods are aggregated into full radiology reports, following a standardized CTPA reporting framework. As a baseline, we also implemented holistic captioning using direct image prompts (e.g., "Describe the abnormal findings in this image."). Variants include organ list guidance and one-shot examples (Supplementary Fig. 1).

Study-level report generation

To standardize the variable response formats of CTPA reading models, we employed a report-writing agent based on the Llama 3 model[22], guided by structured prompts (Supplementary Fig. 2). The agent synthesized paragraph-level image findings into concise, anatomically organized study-

level reports by aggregating information across images, eliminating redundancy, and resolving contradictions.

The output was structured into seven anatomical categories, including Pulmonary Arteries, Lungs and Airways, Pleura, Heart, Mediastinum and Hila, Chest Wall and Lower Neck, and Chest Bones. Only relevant abnormalities were retained per region, and normal regions were explicitly labeled. This approach ensured semantic consistency and localization of findings, producing high-quality reports for downstream tasks such as impression summarization and survival prediction.

To generate the Study Impression section, we employed the report writing agent[22], guided by writing prompts to summarize clinically significant content from the Study Findings (Supplementary Fig. 3). The prompt instructed the model to begin with a definitive conclusion on PE status-for example, explicitly stating "No pulmonary embolism is identified" when applicable-ensuring consistent prioritization of PE-related findings.

The model then distilled the most acute and clinically relevant abnormalities, particularly those related to PE, such as cardiopulmonary, vascular, or parenchymal changes. Normal or less critical findings (e.g., "no acute abnormality") were excluded to enhance focus. The impressions were output in a structured, numbered format with professional, precise language, supporting clarity, clinical prioritization, and downstream decision-making.

Multimodal survival prediction

The Survival Prediction (SP) model estimates patient prognosis by integrating four modalities: CTPA images, generated radiology reports, abnormality identification, and PESI clinical variables. Each modality is processed by a dedicated SP module to estimate modality-specific survival risks, as illustrated in Fig. 2D.

CTPA features are extracted from the multi-abnormality classifier after average pooling. The study findings reports generated by CT-CHAT with 'Abn-Pred' are encoded using a pretrained BERT model[31]. Each SP module comprises a survival encoder (E_s^m) and a risk predictor (C_s^m), which together compute the survival risk for modality m as:

$$R^m = C_s^m(E_s^m(F^m)), \quad m \in [\text{Img}, \text{Text}, \text{Dia}, \text{Var}] \tag{1}$$

For image and text modalities, the survival encoder is a 3-layer MLP with input sizes of 2048 and 4096, and hidden layers of 1024, 512, and 128 nodes with ReLU activations. Diagnosis and clinical features are processed using a 2-layer MLP with 512 and 128 hidden units. Each SP module is optimized using the Cox proportional hazards (CoxPH) loss[32], which is suitable for censored survival data:

$$L_{\text{CoxPH}} = -\frac{1}{N_{Y_e=1}} \sum_{i:y_e^i=1} \left(r^i - \log \sum_{j:y_t^j > y_t^i} e^{r^j} \right) \tag{2}$$

where r^i is the predicted risk score for the i-th patient, y_t^i is the observed survival duration, and y_e^i is the mortality event indicator. This loss function promotes concordance between predicted risk and survival time. Models are trained using the AdamW optimizer with a learning rate of 0.001 for up to 15 epochs, and batch normalization is applied to stabilize training. The final survival prediction is derived by fusing the four modality-specific risk scores using a CoxPH regression model[23]:

$$h_i(t) = h_0(t) \exp\left(\sum_m \beta^m R^m \right) \tag{3}$$

where $h_i(t)$ is the hazard function for patient i at time t, $h_0(t)$ is the baseline hazard, and β^m are the learnable modality-specific coefficients. This fusion strategy allows for effective weighting of each modality's contribution, yielding a robust and accurate survival prediction.

Multimodal data preprocessing

Each CTPA exam was preprocessed by extracting pixel data from DICOM files, standardizing spatial coordinates, and Hounsfield Units (HU). All scans from a single patient session were combined into a unified CTPA image stack, with lung areas segmented and cropped with a 20 mm margin to focus on the chest[33]. These reformatted images were then prepared for abnormality classification and report generation. Axial images were resampled to 1.5 mm in-plane resolution and 3 mm out-of-plane, then padded, and cropped to $224 \times 224 \times 160$ for the abnormality classifier. HU values were normalized to a 0–1 range by clipping values outside the −1000 to 1000 range.

We extracted the "Findings" and "Impression" sections from the original radiology reports as the primary content for our study. We employed the Llama 3[22] model with LLM's medical capabilities to automatically extract 32 anomaly labels from the Findings section, specifying the presence or absence of specific conditions (Supplementary Fig. 4). This automated extraction process improved efficiency over manual annotation and enhanced consistency and accuracy for identifying abnormal findings. The extracted anomaly labels served as reference standards for training and evaluation. The population distributions in two datasets are shown in Fig. 1b.

From the EHR data of the prognosis cohorts, we extracted time-to-event labels for survival analysis, along with 11 PESI variables gathered from clinical data during the retrospective chart reviews. The PESI variables included age, sex, comorbidities (cancer, heart failure, and chronic lung disease), pulse rate, blood pressure, respiratory rate, temperature, mental status, and arterial oxygen saturation at diagnosis [9]. The demographics of the study population with extracted 11 PESI variables are presented in Table 1.

Abnormality identification training

For diagnosis training, the classifier uses a multi-label binary Cross Entropy Loss function. The target labels for the abnormalities are extracted from the findings of CTPA reports. The training process incorporates several data augmentation techniques, including Center Spatial Crop for cropping, Rand Rotate 90 degrees and Rand Flip for rotations and flips, and intensity adjustments through Rand Scale Intensity and Rand Shift Intensity. The training is conducted with a learning rate of 1e-5 with AdamW optimizer, a batch size of 20, and a maximum of 15 epochs.

Report generation metrics

To assess the quality of generated radiology reports, we used standard NLG metrics that evaluate both lexical and semantic alignment with reference texts. BLEU (Bilingual Evaluation Understudy)[34] quantifies fluency and adequacy based on n-gram precision; ROUGE (Recall-Oriented Understudy for Gisting Evaluation)[35] measures content overlap to evaluate summarization quality. METEOR (Metric for Evaluation of Translation with Explicit ORdering)[36] integrates unigram alignment, stemming, synonymy, and word order to assess alignment with ground truth. BERTScore[37] compares contextualized embeddings derived from pre-trained language models to capture semantic similarity beyond surface-level matching. Together, these metrics provide a comprehensive evaluation of linguistic accuracy and clinical relevance.

Statistical analysis

Statistical analysis was performed to assess the significance and robustness of model performance. For classification metrics such as AUROC and sensitivity, statistical comparisons were conducted across each abnormality and regions to evaluate model performance differences.[38] For report generation, we applied bootstrap resampling to estimate 95% confidence intervals for each evaluation metric (BLEU[34], ROUGE[35], METEOR[36], BERTScore[37]), and paired t-tests or Wilcoxon signed-rank tests were used to determine statistical differences between prompting strategies. In survival prediction, C-index comparisons across models were evaluated using bootstrapped intervals, and stratified risk groups were analyzed via Kaplan–Meier survival curves[27] with log-rank testing. These statistical

procedures ensured that observed improvements were both reliable and clinically meaningful.

Data availability

The BUH and JHU datasets used in this study are not publicly available due to institutional privacy agreements. The INSPECT dataset is publicly accessible for non-commercial use under a data use agreement, and available at the following URL: https://som-shahlab.github.io/inspect-website.

Code availability

The source code is publicly available on GitHub at: https://github.com/zzs95/CTPA-Agent. Supplementary Table 1 details the configurations of the pretrained VLMs and LLMs utilized as agent models during inference. These models were employed without additional fine-tuning to perform disease diagnosis and radiology report generation. All experiments were implemented using PyTorch (v2.5) and Transformers (v4.44), and executed on a workstation equipped with two NVIDIA A6000 GPUs (48 GB each). Survival analysis was performed using the `pycox` package and Huggingface libraries.

Received: 31 January 2025; Accepted: 16 June 2025;
Published online: 12 July 2025

References

1. Beckman, M. G., Hooper, W. C., Critchley, S. E. & Ortel, T. L. Venous thromboembolism: a public health concern. *Am. J. Prevent. Med.* **38**, S495–S501 (2010).
2. Lewis, A. E., Gerstein, N. S., Venkataramani, R. & Ramakrishna, H. Evolving management trends and outcomes in catheter management of acute pulmonary embolism. *J. Cardiothorac. Vasc. Anesth.* **36**, 3344–3356 (2022).
3. Leung, A. N. et al. American Thoracic Society documents: an official American Thoracic Society/Society of Thoracic Radiology clinical practice guideline-evaluation of suspected pulmonary embolism in pregnancy. *Radiology* **262**, 635–646 (2012).
4. Tarbox, A. K. & Swaroop, M. Symposium: embolism in the intensive care unit. *Int. J. Crit. Illn. Inj. Sci.* **3**, 69–72 (2013).
5. Alonso-Martínez, J. L., Sánchez, F. A. & Echezarreta, M. U. Delay and misdiagnosis in sub-massive and non-massive acute pulmonary embolism. *Eur. J. Intern. Med.* **21**, 278–282 (2010).
6. Hendriksen, J. M. et al. Clinical characteristics associated with diagnostic delay of pulmonary embolism in primary care: a retrospective observational study. *BMJ Open* **7**, e012789 (2017).
7. Stein, P. D. et al. Multidetector computed tomography for acute pulmonary embolism. *N. Engl. J. Med.* **354**, 2317–2327 (2006).
8. Belohlávek, J., Dytrych, V. & Linhart, A. Pulmonary embolism, part i: epidemiology, risk factors and risk stratification, pathophysiology, clinical presentation, diagnosis and nonthrombotic pulmonary embolism. *Exp. Clin. Cardiol.* **18**, 129 (2013).
9. Aujesky, D. et al. Derivation and validation of a prognostic model for pulmonary embolism. *Am. J. Respir. Crit. Care Med.* **172**, 1041–1046 (2005).
10. Cahan, N. et al. Multimodal fusion models for pulmonary embolism mortality prediction. *Sci. Rep.* **13**, 7544 (2023).
11. Hamamci, I. E. et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834* (2024).
12. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463* (2023).
13. Bai, F., Du, Y., Huang, T., Meng, M. Q.-H. & Zhao, B. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578* (2024).
14. Huang, S.-C. et al. Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis. In Proc. *37th International Conference on Neural Information Processing Systems*. 17742–17772 (NIPS, 2023).
15. Park, J., Kim, S., Yoon, B., Hyun, J. & Choi, K. M4cxr: Exploring multi-task potentials of multi-modal large language models for chest x-ray interpretation. *arXiv preprint arXiv:2408.16213* (2024).
16. Deperrois, N. et al. Radvlm: a multitask conversational vision-language model for radiology. *arXiv preprint arXiv:2502.03333* (2025).
17. MDPI. *Large language models in healthcare and medical domain: A review*, Vol. 11.
18. Shen, Y. et al. Multi-modal large language models in radiology: principles, applications, and potential. *Abdom. Radiol.* **50**, 2745–2757 (2024).
19. *Multimodal healthcare AI: identifying and designing clinically relevant vision-language applications for radiology*.
20. Tan, S. et al. Pulmonary cta reporting: Ajr expert panel narrative review. *Am. J. Roentgenol.* **218**, 396–404 (2022).
21. Bukhari, S. M. A. et al. Clinical and imaging aspects of pulmonary embolism: a primer for radiologists. *Clin. Imaging* **117**, 110328 (2024).
22. Dubey, A. et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
23. Fox, J. & Weisberg, S. Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression* (2002).
24. Huang, S.-C. et al. PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging **3**, 1–9. https://www.nature.com/articles/s41746-020-0266-y.
25. Harrell Jr, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. & Rosati, R. A. Regression modelling strategies for improved prognostic prediction. *Stat. Med.* **3**, 143–152 (1984).
26. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* **26**, 565–574 (2006).
27. Efron, B. Logistic regression, survival analysis, and the Kaplan-Meier curve. *J. Am. Stat. Assoc.* **83**, 414–425 (1988).
28. Blankemeier, L. et al. Merlin: a vision language foundation model for 3d computed tomography. *Res. Sq.* **rs.3**, rs–4546309 (2024).
29. Jain, S. et al. Radgraph: extracting clinical entities and relations from radiology reports. In Proc. *35th International Conference on Neural Information Processing Systems* (NIPS, 2021).
30. Zhong, Z. et al. Pulmonary embolism survival prediction using multimodal learning based on computed tomography angiography and clinical data. *Journal of thoracic imaging* 10–1097 (2025).
31. Yang, X. et al. A large language model for electronic health records. *npj Digi. Med.* **5**, 194 (2022).
32. Katzman, J. L. et al. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 1–12 (2018).
33. Hofmanninger, J. et al. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *Eur. Radiol. Exp.* **4**, 1–13 (2020).
34. *ORANGE: A Method for Evaluating Automatic Evaluation Metrics for Machine Translation* https://www.aclweb.org/anthology/C04-1072 (COLING, 2004).
35. *ROUGE: A Package for Automatic Evaluation of Summaries* https://www.aclweb.org/anthology/W04-1013 (Association for Computational Linguistics, 2004).
36. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments* https://www.aclweb.org/anthology/W05-0909. (Association for Computational Linguistics, 2005).
37. *BERTScore: Evaluating Text Generation with BERT* https://openreview.net/forum?id=SkeHuCVFDr
38. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).

Acknowledgements
This work was partially supported by the Johns Hopkins Department of Radiology's BriteStar Research Award from the Walter & Mary Ciceric family. The IRAT Lab is partially supported by the National Institutes of Health / National Cancer Institute P30CA006973.

Author contributions
Z.Z. conceived and designed the method, wrote the manuscript. Z.Z. and Y.W. did the internal and external data processing, experiment, and analysis. J.W., W.C.H., V.S., L.B., S.K., Z.M., S.C., G.B., S.H.A. and H.B. contributed materials and clinical expertise. I.K., C.T.L., G.B., C.G., M.A., Z.J. and H.B. supervised the work. All authors contributed to the interpretation of the results and editing of the final manuscript. X.F. provided technical support. The authors had access to partial data. Z.Z., V.S., L.B., S.K. and J.Z. directly accessed and verified the internal data, and Y.W. and H.B. verified the external data. All authors accept the final responsibility to submit for publication and take responsibility for the contents of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01807-8.

Correspondence and requests for materials should be addressed to Zhicheng Jiao or Harrison Bai.