

A Multi-Center Benchmark for Abdominal Disease Diagnosis and Report Generation from Non-Contrast CT

Mariam Elbakry¹, Aliaa Sayed Sheha¹, Salma Hassan Tantawy¹[0009-0002-1484-5883], Aya Yassin¹[0000-0001-7761-1166], Concetto Spampinato³[0000-0001-6653-2577], Karim Lekadir⁴[0000-0002-9456-1612], Xiaomeng Li²[0000-0003-1105-8083], and Marawan Elbatel^{2*}[0009-0008-5021-4281]

¹ Ain Shams University, Cairo, Egypt

² The Hong Kong University of Science and Technology, Hong Kong SAR, China

³ University of Catania, Catania, Italy

⁴ Universitat de Barcelona, Barcelona, Spain

Abstract. Multiphasic contrast-enhanced CT (CECT) is widely used for abdominal lesion characterization, yet it carries inherent risks of contrast-induced nephropathy, escalates acquisition burden, and heavily contributes to radiologist workload. To address these challenges, we introduce a novel multi-center benchmark for multi-organ abdominal disease diagnosis and automated radiology report generation, which learns to synthesize contrast-enhanced findings from single-phase non-contrast CT (NCCT). To support this, we curated a large-scale dataset of paired NCCT-CECT studies and their corresponding contrast-enhanced radiology reports from two centers, partitioned into internal sets and an external validation cohort. Under a unified evaluation protocol, we benchmarked five contemporary deep learning architectures encompassing chest-specific, abdomen-specific, and general-purpose multimodal domains. Extensive experiments demonstrate that NCCT retains diagnostic signals, achieving an average multi-organ AUC of 71.1% on the internal cohort and 66.2% on the external cohort, respectively. By releasing this dataset and standardized benchmark publicly, this study aims to catalyze future research into safer, resource-efficient, and globally accessible contrast-free abdominal imaging workflows. Code is available at: <https://github.com/xmed-lab/TriALS-Report>.

Keywords: Non-contrast CT · Abdominal Lesion Characterization · Automated Radiology Report Generation · Multi-centric Validation

1 Introduction

Multiphasic computed tomography (CT) is the primary clinical standard for evaluating abdominal diseases, as the accurate classification of complex pathologies heavily relies on phase-specific enhancement patterns [16,5,18,17,8,19,20].

* Corresponding author: mkfmeibatel@connect.ust.hk

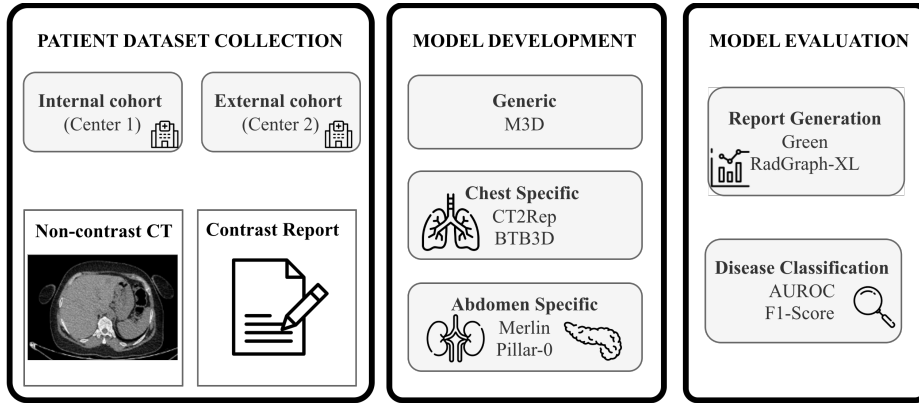


Fig. 1. Overview of the benchmark workflow.

However, the use of iodinated contrast agents poses significant clinical and logistical challenges. Contrast administration is contraindicated in patients with renal impairment or hypersensitivity, and it creates substantial barriers in resource-constrained settings where economic limitations or supply chain instabilities restrict its availability. Non-contrast CT (NCCT) offers a safer, more globally accessible alternative. While the diagnostic features essential for comprehensive abdominal evaluation are often imperceptible to the human eye on NCCT alone, deep learning models have the potential to extract these latent, sub-visual morphological indicators, presenting an opportunity to bypass the need for contrast administration. Therefore, this study aims to evaluate the feasibility of generating comprehensive, contrast-enhanced radiological reports directly from single-phase pelviabdominal NCCT, utilizing ground-truth annotations derived from their contrast-enhanced counterparts.

Despite these opportunities, automated radiology report generation has predominantly been explored in thoracic imaging, with prior work heavily focused on chest X-rays and CT studies [9,10,11,22,13]. In abdominal imaging, deep learning efforts have largely been restricted to isolated diagnostic classification or lesion characterization tasks rather than comprehensive report generation [12]. Furthermore, these existing approaches frequently depend on contrast-enhanced data [3,1,21]. Although recent studies have explored the use of NCCT as an alternative to contrast-enhanced imaging, they have been strictly confined to single-organ or single-pathology applications, such as pancreatic cancer detection, automated urinary stone reporting, or liver lesion segmentation [4,14,7]. Crucially, the advancement of this field is severely hindered by a lack of publicly available data. Consequently, no prior work has introduced a dataset and benchmark enabling multi-organ abdominal lesion detection, characterization, and full report generation from non-contrast CT alone.

To bridge this critical gap, we introduce a novel dataset and comprehensive benchmark specifically designed for the automated generation of contrast-enhanced radiological reports using only non-contrast CT as the input. To the

Table 1. Dataset across the participating centers. Each case consists of a non-contrast CT volume paired with a comprehensive triphasic radiology report.

Feature	Internal	External	Total
Patients / Volumes	1,085	169	1,254
Train Split	760	–	760
Val Split	106	–	106
Test Split	219	169	388
Matrix Size	512 × 512		
In-plane Spacing (mm)	0.875 × 0.875		
Slice Thickness (mm)	1.07		

best of our knowledge, this study is among the first to evaluate the feasibility of mapping the latent features of single-phase pelviabdominal NCCT scans to complex, multi-organ disease characterizations derived from contrast-enhanced reference reports. By establishing a standardized benchmark for this task, our work provides a foundational step toward reducing radiologist workload and improving diagnostic efficiency. Ultimately, successful automated reporting from NCCT alone has the potential to significantly enhance patient safety and expand access to high-quality abdominal diagnostics in environments where multiphasic imaging is unavailable.

2 Methodology

2.1 Overall Benchmark

As illustrated in Fig. 1, our benchmark follows a structured four-stage workflow: (1) Patient Dataset Collection, involving the curation of multi-center NCCT volumes paired with triphasic clinical ground truth; (2) Model Development, where five architectures across three paradigms are fine-tuned on the internal cohort to adapt to the Egyptian population; (3) Model Evaluation, featuring an LLM-based scoring and abdominal disease diagnosis performance. This end-to-end framework ensures that models are evaluated not only on their generative capabilities but on their ability to synthesize triphasic-level diagnostic insights from non-contrast spatial features.

2.2 Dataset

This retrospective study was approved by the Faculty of Medicine, Ain Shams University Research Ethics Committee under protocol FMASU MS 261/2026. All DICOM volumes and accompanying radiology reports were de-identified before model training. Since there is a lack of publicly available triphasic CT datasets tailored for comprehensive abdominal pathology analysis, we curated a multi-center cohort comprising 1,254 patients across two tertiary institutions as depicted in Table 1. This dataset, derived from an Egyptian population, provides

vital regional demographic representation. The cohort was stratified into an internal validation set (Center 1, $n = 1,085$) and an external validation set (Center 2, $n = 169$) to ensure model generalizability across institutional heterogeneities. Center 1 was partitioned into 760 training, 106 validation, and 219 testing cases. To maintain high spatial fidelity for capturing fine-grained pathologies, all volumes possess an axial matrix size of 512×512 , characterized by a mean in-plane pixel spacing of 0.875×0.875 mm and an average slice thickness of 1.07 mm. Ground-truth diagnostic labels were extracted from the triphasic radiology reports using the RATE framework [1]. To ensure clinical reliability and mitigate generative hallucinations, a subset of the extracted labels was audited by board-certified radiologists. The LLM achieved an average clinical relevance score of 90% for disease extraction. Minor terminology variations and spelling discrepancies were manually normalized to ensure a consistent label space for model training.

2.3 Model Development

The evaluated models encompass three distinct archetypes: non-abdominal CT report generation frameworks, generic multimodal 3D architectures, and abdominal CT-specific models. First, we evaluate **CT2Rep** [9], an architecture originally trained on chest CT studies that employs a 3D vision encoder coupled with a novel auto-regressive causal transformer and relational memory. Similarly, we integrate **BTB3D** [10], a causal convolutional encoder-decoder that unifies 2D and 3D training while producing frequency-aware volumetric tokens. Though optimized for the chest, its tokenization approach provides a robust baseline for high-resolution 3D medical imaging. Second, we evaluate **M3D** [2], representing the generic multimodal 3D report generation paradigm. Capable of processing disparate modalities including CT, MRI, and PET, M3D utilizes a 3D spatial pooling perceiver and a large language model backbone, pretrained on multi-organ datasets. Finally, we incorporate domain-specific foundational models: **Merlin** [3], a 3D vision-language model originally pre-trained on large-scale pancreatic and abdominal CT scans paired with radiology reports; and **Pillar-0** [1], a highly sample-efficient radiology foundation model trained across massive datasets of chest, abdomen, and brain volumes. Table 2 summarizes the core technical attributes of these baselines.

2.4 Benchmark Tasks and Model Evaluation

The benchmark formalizes two primary tasks: multi-label disease classification, identifying the presence of distinct abdominal pathologies, and automated report generation, synthesizing a coherent, free-text radiological findings. We standardize the evaluation protocol by decoupling linguistic metrics from clinical efficacy (CE) metrics to prioritize diagnostic performance.

Report Generation Evaluation. Historically, AI-generated radiology reports have been evaluated using standard natural language generation (NLG) metrics such as BLEU, ROUGE, and METEOR. However, contemporary literature

Table 2. Summary of the five baseline deep-learning models, detailing their core modalities, spatial processing dimensions, and original pretraining data domains. *All models have been finetuned on internal training set.*

Model	Modality Focus	2D/3D	Pretraining Dataset Domain
CT2Rep [9]	CT	3D	Chest CT (Diverse populations)
BTB3D [10]	CT	3D	Chest CT (Large-scale corpora)
M3D [2]	Multimodal (CT, MRI, PET)	3D	Multi-organ medical datasets
Merlin [3]	CT	3D	Abdominal & Pancreatic CT
Pillar-0 [1]	Multimodal (CT, MRI)	3D	Chest, Abdomen, Head volumes

demonstrates that these text-overlap metrics are systematically flawed in the medical domain, as they are heavily biased toward matching templated boilerplate text and fail to penalize catastrophic clinical errors like hallucinated findings or reversed negations. To address this, we further incorporate GREEN [15] and RadGraph-XL (Entity, Partial, and Complete) [6] to evaluate structured clinical entity extraction and relation correctness. Traditional NLG metrics are reported strictly as secondary, descriptive endpoints.

Classification Evaluation. For the disease classification task, models are evaluated using the Area Under the Receiver Operating Characteristic curve (AUROC) and the macro-averaged F1-score across all 53 abdominal pathologies. This multi-metric approach ensures a robust, clinical-facing evaluation of AI generalizability in a diverse, non-contrast setting despite the inherent challenges of legacy scanner protocols and retrospective data.

3 Results

Contrast Free Automated Report Generation. The generative capabilities of the models are evaluated in Table 3. Emphasizing clinical correctness over rudimentary text-overlap, we prioritize the GREEN and RadGraph-XL metrics. The zero-shot foundational models (M3D and Merlin) exhibit catastrophic failure when applied out-of-the-box to complex, multi-organ abdominal reporting, severely underperforming across all clinical metrics. Conversely, fine-tuning within our benchmark framework yields massive quantitative gains. We first observe this profound impact on the foundational Merlin architecture: transitioning from zero-shot inference to task-specific fine-tuning catapults its GREEN score on the internal cohort from 6.254 to 34.988, a staggering 459.4% relative improvement. While Merlin’s performance exhibits an expected degradation when exposed to the unseen external Center 2 distribution (achieving a GREEN score of 26.597), it nevertheless maintains a remarkable 435.4% improvement over its external zero-shot baseline (4.967), underscoring the necessity of targeted adaptation for complex abdominal reporting from NCCT compared to Zero-shot.

Figure 2 provides a qualitative comparison between the ground-truth report (derived from the clinical triphasic standard) and the AI-generated report utilizing strictly NCCT inputs. The qualitative alignment powerfully reinforces the

Table 3. Comparison of radiology report generation methods on internal (Center 1) and external (Center 2) datasets for non-contrast reports, evaluated via GREEN [15] and RadGraph-XL [6]. All methods were trained on the internal training set, except for the zero-shot baselines.

Method	BLEU-4	METEOR	BERT	ROUGE-L	GREEN	RadGraph-XL		
						E	P	C
Internal Validation								
M3D (Zero-Shot)	0.056	6.062	78.451	8.857	1.695	4.754	3.857	1.868
Merlin (Zero-Shot)	0.537	13.451	78.513	13.515	6.254	19.859	15.154	12.165
M3D	16.754	31.298	85.108	32.402	35.626	36.328	32.251	26.286
Merlin	21.944	38.843	85.291	32.351	34.988	41.517	36.797	30.190
CT2Rep	17.739	41.130	87.594	49.334	64.978	52.895	46.118	44.139
BTB3D	27.796	39.183	85.970	36.137	36.734	43.451	38.549	32.079
External Validation								
M3D (Zero-Shot)	0.072	5.952	78.116	8.527	1.279	4.944	3.970	1.460
Merlin (Zero-Shot)	0.594	17.000	78.543	13.732	4.967	19.398	14.358	11.261
M3D	6.160	21.668	81.806	21.301	24.192	23.021	19.893	13.896
CT2Rep	2.617	26.114	83.460	25.004	40.545	29.196	25.264	22.289
Merlin	8.731	28.822	82.766	22.245	26.597	27.585	23.745	16.905
BTB3D	6.920	26.813	82.702	21.983	29.132	27.495	23.437	16.349

quantitative gains observed in Table 3. The AI framework successfully synthesizes contrast-enhanced findings, accurately identifying diffuse fatty infiltration of the liver, normal portal vein patency without thrombosis, and subtle, localized findings such as a small left renal cortical cyst. Furthermore, the model accurately details regional inflammatory markers, correctly describing mesenteric fat stranding and haziness consistent with mesenteric panniculitis. By successfully extracting these sub-visual morphological indicators from a single-phase scan, the qualitative results confirm the framework’s potential to generate highly accurate, comprehensive radiological reports while entirely bypassing the clinical risks and logistical burdens of contrast administration.

Disease Diagnosis Performance. The diagnostic classification capabilities of the evaluated architectures across 15 organs and 53 distinct abdominal pathologies are detailed in Table 4. As anticipated, models utilizing multiphasic contrast-enhanced inputs establish the upper performance bound, reflecting the rich temporal enhancement dynamics inherent to the clinical gold standard. Merlin achieves the highest overall diagnostic performance on multiphasic data, securing an average AUC of 76.7% and 72.2% on the internal and external cohorts, respectively. Crucially, when restricted to single-phase NCCT inputs, the models demonstrate remarkable diagnostic retention and highly competitive zero-contrast classification capabilities. Across the 15-organ average, Merlin’s NCCT performance (AUC 71.1%) retains over 90% of its multiphasic diagnostic power (AUC 76.7%). We observe similarly robust retention at the organ-specific level; for instance, Pillar achieves an impressive AUC of 80.4% for pancreatic lesion

Table 4. Disease diagnosis performance across anatomical organs and imaging protocols, reported as percentages with 95% bootstrap confidence intervals (1,000 iterations, patient-level resampling). Multiphase indicates the late fusion of different CT phases. The averages cover 15 organs and a real-world distribution of 53 abdomen CT pathologies from the RATE-Evals [1] taxonomy, including both benign (e.g., simple cysts, hemangiomas) and malignant (e.g., hepatocellular carcinoma, adenocarcinomas) lesions. Best results per organ–centre–protocol–metric are highlighted in bold. Random refers to chance-level discrimination; the F1 random baseline corresponds to predicting each disease at its observed prevalence (macro-averaged across the 15 organs).

Organ / Model	Imaging Protocol	Center 1 Cases		Center 2 Cases	
		AUC (%)	F1 (%)	AUC (%)	F1 (%)
Liver					
Merlin[3]	NCCT	69.48 [66.5,72.4]	34.12 [30.1,37.9]	68.15 [64.3,72.1]	30.79 [27.8,34.0]
	Multiphase	74.39 [71.8,76.9]	38.51 [35.4,41.8]	74.75 [70.9,78.5]	36.66 [34.0,39.3]
Pillar[1]	NCCT	68.24 [65.2,71.5]	24.41 [21.1,27.7]	66.48 [62.4,70.8]	29.33 [25.8,33.5]
	Multiphase	74.55 [71.7,77.4]	32.11 [28.8,35.6]	68.68 [64.5,72.7]	29.23 [26.7,31.8]
M3D[2]	NCCT	62.87 [59.3,66.4]	28.10 [23.8,32.5]	65.28 [61.5,69.0]	29.03 [25.6,32.7]
	Multiphase	67.21 [63.9,70.4]	32.86 [29.6,36.5]	68.69 [64.5,72.3]	33.00 [30.3,35.8]
Spleen					
Merlin[3]	NCCT	61.62 [51.2,72.1]	28.18 [18.7,37.2]	64.96 [54.7,74.8]	34.05 [30.7,37.0]
	Multiphase	62.85 [54.2,71.4]	24.62 [17.2,32.2]	74.44 [66.9,81.8]	40.65 [37.4,43.7]
Pillar[1]	NCCT	65.40 [55.7,74.9]	26.81 [17.5,36.9]	70.47 [63.3,77.4]	30.17 [24.2,35.8]
	Multiphase	61.47 [52.0,72.0]	25.28 [15.2,37.5]	61.81 [54.8,69.8]	34.96 [30.0,39.5]
M3D[2]	NCCT	56.48 [45.4,66.8]	16.93 [7.7,27.3]	60.85 [51.3,70.3]	30.50 [23.9,38.6]
	Multiphase	71.60 [59.6,82.6]	29.91 [21.7,38.4]	59.56 [52.2,67.6]	31.77 [27.2,38.0]
Pancreas					
Merlin[3]	NCCT	76.00 [65.1,85.6]	5.57 [0.0,12.8]	61.32 [52.5,70.1]	17.34 [3.9,32.4]
	Multiphase	87.65 [80.3,94.8]	21.55 [15.8,27.4]	86.13 [77.5,93.1]	14.04 [8.9,19.8]
Pillar[1]	NCCT	80.39 [71.4,87.9]	13.52 [8.8,18.0]	78.55 [64.8,92.1]	8.19 [4.0,11.9]
	Multiphase	83.68 [75.6,90.0]	16.95 [10.1,23.8]	92.84 [88.6,96.3]	11.65 [10.2,13.4]
M3D[2]	NCCT	72.91 [66.2,80.2]	16.28 [0.0,33.3]	56.03 [47.8,63.7]	0.00 [0.0,0.0]
	Multiphase	69.94 [57.6,82.0]	10.45 [6.1,16.6]	69.54 [59.8,79.1]	2.58 [0.0,6.1]
Kidneys					
Merlin[3]	NCCT	58.60 [50.6,66.5]	22.90 [17.9,28.0]	61.00 [53.0,69.3]	20.42 [14.1,27.0]
	Multiphase	60.73 [53.8,67.6]	25.36 [21.0,29.9]	62.28 [54.5,69.8]	21.50 [16.8,26.7]
Pillar[1]	NCCT	63.69 [56.4,70.8]	21.75 [17.4,26.4]	68.78 [61.8,75.4]	17.25 [11.7,23.4]
	Multiphase	75.85 [69.9,81.6]	31.07 [26.9,35.1]	72.11 [62.8,81.1]	29.28 [23.8,34.8]
M3D[2]	NCCT	53.08 [46.4,59.8]	14.70 [10.7,18.8]	49.96 [42.1,57.8]	17.87 [12.3,23.3]
	Multiphase	62.60 [55.3,68.8]	23.87 [18.6,29.8]	56.41 [45.3,67.8]	21.92 [16.5,27.4]
Average across 15 different organs and 53 diseases					
Random (chance) —		50.00	7.52	50.00	9.39
Merlin[3]	NCCT	71.05 [68.8,73.3]	24.97 [22.7,27.4]	66.24 [63.7,68.8]	25.18 [22.8,27.6]
	Multiphase	76.72 [74.6,78.7]	29.17 [27.4,31.0]	72.22 [69.7,74.6]	25.56 [24.2,26.9]
Pillar[1]	NCCT	69.92 [67.4,72.2]	21.74 [19.8,23.6]	65.71 [63.2,68.4]	20.83 [18.8,22.7]
	Multiphase	75.82 [73.9,77.8]	24.55 [22.5,26.5]	70.30 [67.9,72.5]	24.59 [22.9,26.2]
M3D[2]	NCCT	67.91 [65.3,70.5]	21.38 [18.8,24.2]	61.33 [58.8,63.9]	19.96 [18.1,22.1]
	Multiphase	72.42 [70.0,74.8]	24.68 [22.6,26.7]	64.94 [62.2,67.5]	24.14 [22.4,25.9]

Radiologist written using Triphasic CT-Gold Standard	AI generated Report using Non-contrast CT
• The liver is diffusely enlarged measuring 19.8cm at maximum CC dimension, displaying a homogenously reduced CT density suggestive of underlying fatty infiltration, with no definite focal lesion or dilatation of intra hepatic biliary radicles noted. • The portal vein is patent with no thrombosis. • The gall bladder shows normal wall thickness with no stone seen inside. • Normal CT appearance of the spleen, pancreas and both suprarenal glands. • Both kidneys are of normal site, size and parenchymal thickness with no stones or backpressure changes. • A rather well circumscribed area of mesenteric fat stranding and haziness is seen along the root of mesentery, displaying a (misty attenuation appearance), with prominent locoregional mesenteric vascular branches as well as <i>subcentimetric</i> mesenteric lymphadenopathy, mostly consistent with <i>mesenteric panniculitis</i>, for clinical correlation and follow-up. • Normal appearance of the stomach with no sizable soft tissue masses. • A fat-density ovoid structure is seen abutting the anterolateral surface of the distal descending colon, roughly measuring 2x14mm at maximum axial dimensions, exhibiting a central relatively dense linear structure with a surrounding hyperdense rim as well as extensive surrounding fatty stranding and haziness, mostly consistent with <i>epiploic appendagitis</i>, yet with no sizable related encysted or free fluid collection seen, for clinical correlation. • Otherwise, the rest of colonic bowel loops show a normal CT appearance with no evidence of wall thickening, stricture, pericolic fat stranding or mass formation. • Normal filling of the urinary bladder with normal wall thickness and no calc. • The prostate and both seminal vesicles are of normal C.T. appearance.	• The liver is enlarged in size displaying a diffuse decreased attenuation pattern, denoting diffuse fatty infiltration with no evidence of definite focal lesions or dilatation of intra-hepatic biliary radicles. • The portal vein is of average diameter with normal contrast opacification and no CT evidence of internal thrombosis. • The gall bladder shows normal wall thickness with no stone seen inside. • The spleen is of average size displaying a normal and homogenous CT appearance. The pancreas and both adrenal glands show normal C.T. appearance. • Both kidneys are seen in their normal anatomical site displaying average <i>normal</i> and normal shape with no backpressure changes. Both show adequate contrast excretion. A small left renal cortical cyst is noted. • A rather well circumscribed area of mesenteric fat stranding and haziness is seen along the root of mesentery, displaying a (<i>misty attenuation appearance</i>), mostly consistent with mesenteric panniculitis, for clinical correlation and onco-up, for clinical correlation. • Normal CT appearance of the scanned bowel loops, with no evidence of wall thickening, stricture, pericolic fat stranding, or mass formation. • The urinary bladder shows normal wall thickness with no calculi, masses, or diverticulae. • The prostate and both seminal vesicles are of normal C.T. appearance. • Normal ischiorectal fossae on both sides. • No evidence of ascites or pelvic fluid collection. • The lower lung cuts are unremarkable.

Fig. 2. A comparison between AI-generated findings and expert ground truth (GT) using only non-contrast CT (NCCT) inputs.

detection on NCCT (Center 1), effectively bridging the gap to complex triphasic detection rates. Furthermore, Merlin achieves a robust AUC of 69.5% for liver pathologies from NCCT inputs, compared to 74.4% using multiphasic data. This high retention rate across a diverse taxonomy provides compelling empirical evidence that deep learning frameworks can successfully map latent non-contrast features to complex, multi-organ disease characterizations, validating the clinical feasibility of NCCT-driven diagnostics.

4 Discussion

This study introduces the first multi-centric dataset and benchmark dedicated to contrast-free abdominal disease diagnosis and automated report generation. By successfully mapping latent non-contrast spatial features to complex triphasic-equivalent findings, our benchmark proves that AI-driven NCCT retains 90% of AI-driven CECT diagnostic power. Furthermore, the massive quantitative leaps achieved over zero-shot foundational models decisively validate the necessity of our targeted, multi-centric adaptation pipeline for complex abdominal reporting.

However, these findings exist within specific foundational boundaries. First, while performance naturally shifts on the external Center 2 cohort due to scanner variations, this deliberately establishes a realistic, rigorous baseline that invites targeted domain adaptation research. Second, although our scalable framework relies on global text-alignment rather than pixel-level lesion grounding, this in-

tentional design enabled massive multi-centric curation, with expert radiologist audits successfully mitigating label noise to ensure high clinical validity.

Finally, NCCT imposes a fundamental information ceiling. Enhancement-kinetic findings such as arterial hyperenhancement, portal-venous washout, delayed retention, and capsule appearance depend on intravenous contrast dynamics and cannot be physically inferred from a single-phase non-contrast scan. Within this constraint, our benchmark shows that latent morphological signatures recover the majority of AI-driven multiphasic diagnostic performance, with the residual gap concentrated on low-contrast targets such as the pancreas and kidneys, where enhancement timing is most informative. We therefore position NCCT-driven reporting as a complementary triage and screening modality, particularly in settings where contrast administration is contraindicated or unavailable, rather than a replacement for multiphasic protocols. Downstream contrast-enhanced imaging remains warranted for definitive lesion characterization and for findings whose diagnosis depends on enhancement kinetics. We anticipate that the public release of this dataset and benchmark will catalyze future research into contrast-free imaging, ultimately accelerating the routine clinical deployment of safer, more globally accessible diagnostic workflows.

Acknowledgments. The work has been partially supported by the European Union – Next Generation EU, Mission 4 Component 2 Line 1.3, through the PNRR MUR project PE0000013 – FAIR “Future Artificial Intelligence Research” (CUP E63C22 001940006). Marawan Elbatel is supported by the Hong Kong PhD Fellowship Scheme from the Hong Kong Research Grants Council (RGC), and partially supported by a research grant from the Bridge Gap Fund (Project No. BGF.020.2025) at HKUST.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agrawal, K.K., Liu, L., Lian, L., Nercessian, M., Harguindeguy, N., Wu, Y., Mikhael, P., Lin, G., Sequist, L.V., Fintelmann, F., Darrell, T., Bai, Y., Chung, M., Yala, A.: Pillar-0: A new frontier for radiology foundation models. arXiv preprint arXiv:2511.17803 (2025)
2. 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 **abs/2404.00578** (2025)
3. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truys, C., et al.: Merlin: A vision language foundation model for 3d computed tomography. Research Square pp. rs-3 (2024)
4. Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., Tang, W., Jin, G., Jiang, H., Fang, X., Nogues, I., Li, X., Guo, W., Wang, Y., Fang, W., Qiu, M., Hou, Y., Kovarnik, T., Vocka, M., Lu, Y., Chen, Y., Chen, X., Liu, Z., Zhou, J., Xie, C., Zhang, R., Lu, H., Hager, G.D., Yuille, A.L., Lu, L., Shao, C., Shi, Y., Zhang, Q., Liang, T., Zhang, L., Lu, J.: Large-scale pancreatic cancer detection via non-contrast ct and deep learning. Nature Medicine

- 29**(12), 3033–3043 (Nov 2023). <https://doi.org/10.1038/s41591-023-02640-w>, <http://dx.doi.org/10.1038/s41591-023-02640-w>
5. Chernyak, V., Fowler, K.J., Kamaya, A., Kielar, A.Z., Elsayes, K.M., Bashir, M.R., Kono, Y., Do, R.K., Mitchell, D.G., Singal, A.G., Tang, A., Sirlin, C.B.: Liver imaging reporting and data system (li-rads) version 2018: Imaging of hepatocellular carcinoma in at-risk patients. *Radiology* **289**(3), 816–830 (Dec 2018)
 6. Delbrouck, J.B., Chambon, P., Chen, Z., Varma, M., Johnston, A., Blankemeier, L., Van Veen, D., Bui, T., Truong, S., Langlotz, C.: RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) *Findings of the Association for Computational Linguistics: ACL 2024*. pp. 12902–12915. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024)
 7. Elbatel, M., Ghonim, M., Mao, J., Lin, Z., et al.: Trials: Triphasic-aided liver lesion segmentation benchmark in non-contrast ct (2026), <https://arxiv.org/abs/2605.16572>
 8. Gao, R., Zhao, S., Aishanjiang, K., Cai, H., Wei, T., Zhang, Y., Liu, Z., Zhou, J., Han, B., Wang, J., Ding, H., Liu, Y., Xu, X., Yu, Z., Gu, J.: Deep learning for differential diagnosis of malignant hepatic tumors based on multi-phase contrast-enhanced ct and clinical data. *Journal of Hematology & Oncology* **14**(1) (Sep 2021). <https://doi.org/10.1186/s13045-021-01167-2>, <http://dx.doi.org/10.1186/s13045-021-01167-2>
 9. Hamamci, I.E., Er, S., Menze, B.: CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging . In: *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. vol. LNCS 15012. Springer Nature Switzerland (October 2024)
 10. Hamamci, I.E., Er, S., Shit, S., Reynaud, H., Yang, D., Guo, P., Edgar, M., Xu, D., Kainz, B., Menze, B.: Better tokens for better 3d: Advancing vision-language modeling in 3d medical imaging. In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems* (2025)
 11. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Durugol, O.F., Hou, B., Shit, S., Dai, W., Xu, M., Reynaud, H., Dasdelen, M.F., Wittmann, B., Amiranashvili, T., Simsar, E., Simsar, M., Erdemir, E.B., Alanbay, A., Sekuboyina, A., Lafci, B., Kaplan, A., Lu, Z., Polacin, M., Kainz, B., Bluethgen, C., Batmanghelich, K., Ozdemir, M.K., Menze, B.: Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering* (Feb 2026)
 12. Jorg, T., Halfmann, M.C., Stoeck, F., Arnhold, G., Theobald, A., Mildnerberger, P., Müller, L.: A novel reporting workflow for automated integration of artificial intelligence results into structured radiology reports. *Insights into Imaging* **15**(1) (Mar 2024). <https://doi.org/10.1186/s13244-024-01660-5>, <http://dx.doi.org/10.1186/s13244-024-01660-5>
 13. Lee, J.O., Cho, J., Kim, J., Dillard, L., van Sonsbeek, T., Setio, A.A.A., Lee, H., Yoo, D., Kim, T.: SPEC-CXR: Advancing Clinical Safety through Entity-Level Performance Evaluation of Chest X-ray Report Generation . In: *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. vol. LNCS 15966. Springer Nature Switzerland (September 2025)
 14. Liang, R., Fan, Z., Pan, J., Sun, C., Terry, R.S., Xu, J.: A clinically-grounded two-stage framework for renal ct report generation. *ArXiv abs/2506.23584* (2025)
 15. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Michalson, A.E., Moseley, M., Langlotz, C., Chaudhari, A.S., Delbrouck, J.B.: GREEN:

- Generative radiology report evaluation and error notation. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 374–390. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024)
16. Rydberg, J., Buckwalter, K.A., Caldemeyer, K.S., Phillips, M.D., Conces, D.J., Aisen, A.M., Persohn, S.A., Kopecky, K.K.: Multisection ct: Scanning techniques and clinical applications. *RadioGraphics* **20**(6), 1787–1806 (Nov 2000)
 17. Shan, R., Pei, C., Fan, Q., Liu, J., Wang, D., Yang, S., Wang, X.: Artificial intelligence-assisted platform performs high detection ability of hepatocellular carcinoma in ct images: an external clinical validation study. *BMC Cancer* **25**(1) (Jan 2025). <https://doi.org/10.1186/s12885-025-13529-x>, <http://dx.doi.org/10.1186/s12885-025-13529-x>
 18. Shan, Z., Zhang, X., Zhang, Y., Wang, S., Wang, J., Shi, X., Li, L., Li, Z., Yang, L., Liu, H., Li, W., Yang, J., Yang, L.: Combined nomogram for differentiating adrenal pheochromocytoma from large-diameter lipid-poor adenoma using multi-phase ct radiomics and clinico-radiological features. *BMC Medical Imaging* **25**(1) (Aug 2025)
 19. Uhm, K.H., Jung, S.W., Choi, M.H., Shin, H.K., Yoo, J.I., Oh, S.W., Kim, J.Y., Kim, H.G., Lee, Y.J., Youn, S.Y., Hong, S.H., Ko, S.J.: Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography. *npj Precision Oncology* **5**(1) (Jun 2021)
 20. Wu, X., Su, H., Hua, Y., Xu, Y., Wang, L., Wang, X., Wang, S., Jin, B., Liu, X., Wan, X., Sun, Q., Wang, X., Du, S.: A multi-phase ct dataset for automated differential diagnosis of liver tumors. *Scientific Data* **13**(1) (Dec 2025)
 21. Yamamoto, K., Kikuchi, T.: Totalfm: An organ-separated framework for 3d-ct vision foundation models. *ArXiv abs/2601.00260* (2026)
 22. Zhang, B., Jia, C., Liu, S., Schunkert, H., Navab, N.: Semantic-Aware Chest X-ray Report Generation with Domain-Specific Lexicon and Diversity-Controlled Retrieval. In: *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. vol. LNCS 15965. Springer Nature Switzerland (September 2025)