This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

UniCT: A Unified Joint Multi-Task Framework for 3D Chest CT Abnormality Analysis

Theo Di Piazza[1,2], Carole Lazarus[3], Olivier Nempont[3], and Loic Boussel[1,2]

[1]INSA Lyon, University of Lyon, CNRS, INSERM, CREATIS UMR 5220, U1294
[2]Hospices Civil de Lyon
[3]Philips Clinical Informatics
theo.dipiazza@creatis.insa-lyon.fr

Abstract. The increasing number of CT scan examinations has lead to the need of developing automated tools to support radiologists managing their growing workload. Multi-label abnormality analysis from 3D chest CT scans, including abnormality classification, segmentation and report generation, is therefore a key capability for clinical decision support. However, such tasks remain highly challenging due to the high-dimensional structure of the data, and the wide variety of abnormalities to detect. Existing approaches typically address abnormality classification, segmentation, or report generation in isolation, despite their strong semantic and clinical interdependence. In clinical workflows, radiologists identify abnormalities by simultaneously localizing regions of interest and synthesizing descriptive findings. In light of this, we propose UniCT, the first end-to-end unified framework that jointly performs multi-label abnormality classification, segmentation, and report generation from 3D chest CT scans. UniCT explicitly models cross-task interactions through a multi-task fusion mechanism and a segmentation-conditioned feature modulation, enabling shared spatial and semantic reasoning across tasks. Experimental results on two public datasets demonstrate that joint learning yields consistent mutual gains across tasks, achieving competitive performance and providing an effective inductive bias for CT analysis.

Keywords: Multi-Task Learning · Chest Computed Tomography · Report Generation · Multi-label Abnormality Classification · Segmentation.

1 Introduction

Computed Tomography (CT) provides detailed volumetric visualization of the human body and plays a central role in the detection and characterization of thoracic abnormalities [27]. However, the increasing number of examinations results in the need to develop automated tools to assist radiologists managing such a growing workload [6]. In this context, multi-label abnormality analysis has emerged as a key paradigm, supporting tasks such as abnormality classification [13,20], segmentation [2,15,9], and automated report generation [17,24,8].

Prior work shows that modeling multiple abnormalities improves diagnostic performance by capturing inter-abnormality dependencies [13] and provides

strong visual representations for downstream tasks such as report generation [11]. Nevertheless, multi-label abnormality analysis in 3D Chest CT remains challenging due to the volumetric complexity of the data and heterogeneous abnormality appearance [16]. Recent advances in representation learning architectures [12,33,22,19] and large-scale CT foundation models [16,26,5,35] have substantially improved performance across individual downstream tasks. Despite these advances, most existing methods address classification, segmentation and report generation independently, limiting their ability to exploit complementary spatial and semantic information across tasks. Classification models often lack spatial interpretability, while report generation systems are highly sensitive to the quality of pretrained visual features [3], and often rely on large-scale pretraining [5] or multi-stage pipelines [11], requiring sequential optimization which restricts joint reasoning across prediction, localization, and description.

Joint multi-task learning that explicitly integrates abnormality classification, segmentation and report generation remains underexplored in 3D CT, largely due to the scarcity of datasets providing aligned reports, labels, and segmentations. The recent release of `ReXGroundingCT` [2] partially addresses this limitation by pairing textual descriptions with abnormality segmentations, enabling grounded report generation [4], text-prompted medical segmentation [14,39], and opening new opportunities for end-to-end multi-task learning, which we investigate in this work. While prior works have explored multi-task settings in 2D imaging [32], existing methods for 3D CT typically focus on single abnormalities and do not incorporate report generation [38,23], limiting their ability to leverage multi-label clinical reasoning and the semantic supervision provided by language modeling, further motivating our proposed unified framework.

In clinical practice, radiologists simultaneously localize abnormalities, assess their clinical significance, and synthesize coherent textual reports by leveraging shared spatial and semantic cues. This observation suggests that abnormality classification, segmentation, and report generation could benefit from joint modeling, rather than isolated optimization. Motivated by this clinical workflow, our work introduces UniCT, the first end-to-end unified approach to jointly perform multi-label abnormality classification, segmentation, and report generation from 3D CT. UniCT incorporates a multi-task fusion module that enable information exchange across classification, segmentation, and report generation branches, allowing the model to leverage spatial localization signals to improve descriptive reporting and contextual semantic cues to refine abnormality recognition.

Contributions. (1) We introduce UniCT, a unified end-to-end framework that jointly performs abnormality classification, segmentation, and report generation from 3D chest CT volumes. **(2)** We propose a multi-task interaction mechanism combining bidirectional task fusion with segmentation-conditioned feature modulation, enabling shared spatial reasoning across tasks. **(3)** Through experiments on two public datasets, we show that unified multi-task learning consistently improves performance across abnormality classification, segmentation, and report generation compared with task-specific baselines.

2 Method

As shown in Fig. 1, UniCT processes a 3D CT scan through a visual encoder to obtain slice-level and global representations of the volume (Sec. 2.1). Task-specific latent representations for abnormality classification, segmentation, and report generation are then derived via adapters and refined through a multi-task fusion module to enable information exchanges across tasks (Sec. 2.2). Features are finally passed to independent modules for task-specific predictions (Sec. 2.3).

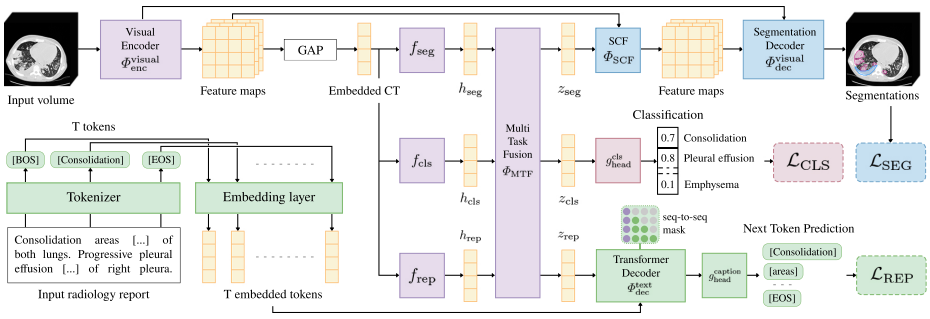

Fig. 1: UniCT jointly addresses abnormality classification, segmentation and report generation in a unified approach, complemented by a multi-task fusion and a segmentation-conditioned feature module to enable cross-task interactions.

2.1 Visual Features Extraction

Motivated by the empirical success of 2.5D modeling over 3D encoders for end-to-end chest CT analysis [13,12], the CT scan $X \in \mathbb{R}^{C\times H\times W}$ is represented as a set of $S = 80$ non-overlapping axial slice triplets, with $(C, H, W) = (240, 480, 480)$. A 2D ResNet18 $\Phi_{\text{enc}}^{\text{visual}}$, shared across all tasks, encodes each triplet into volumetric feature maps $H_{\text{CT}} \in \mathbb{R}^{S\times h\times w\times d}$ with $d = 512$ and $h = w = 15$. This 2.5D formulation offers a favorable trade-off between representational capacity and computational efficiency, enabling effective end-to-end multi-task training without relying on CT-domain-specific pretraining. For global prediction and report generation, slice-level features are aggregated via Global Average Pooling (GAP) to obtain a compact scan-level representation $h_{\text{CT}} \in \mathbb{R}^d$, while spatial feature maps are preserved for dense prediction:

$$h_{\text{CT}} = f_{\text{GAP}}(H_{\text{CT}}) = \left(f_{\text{GAP}} \circ \Phi_{\text{enc}}^{\text{visual}}\right)(X)\,. \tag{1}$$

2.2 Task-Specific Latent Representations

Task-specific Adapters. To balance representation sharing and task specialization, h_{CT} is projected into three task-specific d-dimensional latent spaces using lightweight multi-layer perceptron (MLP) adapters. Given the heterogeneous

supervision signals across abnormality classification (scan-level), segmentation (voxel-level), and report generation (sequence-level), these adapters encourage the emergence of task-specific subspaces. The adapters f_{seg}, f_{cls} and f_{rep}, consist of a linear projection followed by a ReLU activation, producing task-specific embeddings defined as $h_{\text{cls}} = f_{\text{cls}}(h_{\text{CT}})$, $h_{\text{seg}} = f_{\text{seg}}(h_{\text{CT}})$ and $h_{\text{rep}} = f_{\text{rep}}(h_{\text{CT}})$.

Multi-Task Fusion Module. While task-specific adapters enable specialization, clinical reasoning inherently requires consistency across global diagnosis, localization and textual description. To explicitly model these cross-task dependencies, we introduce a Multi-Task Fusion (MTF) module Φ_{MTF}, implemented as a Transformer encoder operating over the set of task embeddings. Through self-attention [34], MTF adaptively reweights and exchanges information across task representations, allowing, for instance, segmentation-derived spatial cues to inform report generation or classification priors to regularize dense predictions:

$$\{z_{\text{cls}}, z_{\text{seg}}, z_{\text{rep}}\} = \Phi_{\text{MTF}}(h_{\text{cls}}, h_{\text{seg}}, h_{\text{rep}})\,. \tag{2}$$

2.3 Task-Specific Objectives

Abnormality Classification. The refined classification embedding z_{cls} is processed by a classification head $g^{\text{cls}}_{\text{head}}$ to predict logits $\hat{y} \in \mathbb{R}^M$ over M abnormalities. We apply a Binary Cross Entropy loss $\mathcal{L}_{\text{CLS}}$ on the predicted logits against the ground-truth multi-label annotations $y \in \{0,1\}^M$.

Segmentation. To enable dense prediction while preserving cross-task consistency, UniCT incorporates a Segmentation-Conditioned Feature modulation (SCF) [28], enabling conditioning of spatial CT features using the task-refined semantic embedding z_{seg}, allowing abnormality cues to guide dense predictions:

$$Z_{\text{CT}} = \Phi_{\text{SCF}}(H_{\text{CT}}, z_{\text{seg}}) = \gamma(z_{\text{seg}}) \odot H_{\text{CT}} + \beta(z_{\text{seg}})\,, \tag{3}$$

where $\odot$ denotes the element-wise multiplication, while γ and β represent the scaling and shift affine transformation (linear projection with ReLU). The modulated features $Z_{\text{CT}} \in \mathbb{R}^{S\times h\times w\times d}$ are processed by a segmentation decoder $\Phi^{\text{visual}}_{\text{dec}}$ enhanced with skip connections [29], to produce abnormality-specific segmentation masks $\hat{X} \in \mathbb{R}^{M\times C\times H\times W}$, defined as $\hat{X} = \Phi^{\text{visual}}_{\text{dec}}(Z_{\text{CT}})$. We supervise segmentation using a Dice loss, denoted as $\mathcal{L}_{\text{SEG}}$ [31].

Report Generation. The radiology report R is tokenized and embedded into $H_{\text{rep}} \in \mathbb{R}^{T\times d}$. The T textual tokens are decoded autoregressively by a Transformer decoder $\Phi^{\text{text}}_{\text{dec}}$, conditioned on the cross-task refined representation z_{rep}, where a causal attention mask ensures that each token attends only to preceding tokens and z_{rep}. The predicted next tokens $\hat{R}$ are obtained via a captioning head $g^{\text{caption}}_{\text{head}}$, optimizing a token-level Cross-Entropy loss $\mathcal{L}_{\text{REP}}$, such that:

$$H_{\text{rep}} = \left(\Phi^{\text{text}}_{\text{emb}} \circ \Phi^{\text{text}}_{\text{token}}\right)(R), \quad \text{and} \quad \hat{R} = \left(g^{\text{caption}}_{\text{head}} \circ \Phi^{\text{text}}_{\text{dec}}\right)(H_{\text{rep}}, z_{\text{rep}})\,. \tag{4}$$

Objective. UniCT is trained by minimizing the sum of all task losses, defined as $\mathcal{L} = \mathcal{L}_{\text{CLS}} + \mathcal{L}_{\text{SEG}} + \mathcal{L}_{\text{REP}}$. All losses are equally weighted, as we empirically observed stable convergence without additional task-balancing strategies.

3 Experiments

Datasets. We use `ReXGroundingCT` [2], a subset of `CT-RATE` [16] containing non-contrast chest CTs from 3,142 patients. From 14 *focal* and *non-focal* abnormalities with segmentation annotations, we retain 10, excluding two umbrella (*other*) and two very low prevalence (< 20 cases) classes to ensure reliable cross-validation. Experiments use 5-fold cross-validation patient-level splits (70/15/15). To assess generalization, we evaluate on the 1,344 patients from the `RAD-ChestCT` [13] validation and test sets, which provide label annotations but no report or segmentation, focusing on the 9 abnormalities that can be mapped (*pulmonary nodes* and *micronodules* combined as *nodes*). CT scans are reformatted to SLP orientation, spacing of $1.5 \times 0.75 \times 0.75$ mm, and center-cropped or padded to a size of $240 \times 480 \times 480$. Hounsfield Units are clipped to $[-1000, 200]$ reflecting practical diagnostic limits [10], rescaled to $[0, 1]$ and normalized with ImageNet statistic [30]. **Implementation.** Training runs for 20,000 iterations using AdamW, $(\beta_1, \beta_2) = (0.9, 0.99)$ and a batch size of 4. To facilitate simultaneous convergence across tasks, the learning rate is set at 5×10^{-5} for the segmentation modules and 5×10^{-6} for the other parameters. For all methods, visual encoders are initialized from ImageNet pretrained weights, either by direct initialization or weight inflation [7], and all parameters are trainable to ensure fair comparison. The training duration for UniCT is 8 hours and inference takes approximately 150 milliseconds on a single NVIDIA RTX A6000 GPU.

3.1 Multi-label Abnormality Classification

Protocol. We first evaluate UniCT on multi-label abnormality classification. UniCT is compared against 2.5D & 3D visual encoders [7,1,13,12], each trained in a single-task settings. To the best of our knowledge, no prior work jointly addresses multi-label abnormality classification, segmentation and report generation from 3D CT within a unified end-to-end framework. To enable a rigorous evaluation, we therefore construct a multi-task baseline (*Multi-task base.*) that shares the same backbone architecture and supervision signals of UniCT, while excluding its task-interaction components (adapters, MTF and SCF). This design isolates the contribution of UniCT's unified objective and cross-task reasoning mechanisms, providing a principled benchmark for unified CT analysis.

Quantitative analysis. Table 1 shows that UniCT consistently outperforms both classification-only (■) and the multi-task (⧫) baselines on `ReXGroundingCT` across all evaluation metrics, with statistically significant improvements (paired t-test across folds, $p < 0.01$). Our model demonstrates an overall F1-Score gain of +12% on *focal* and +24% on *non-focal* abnormalities, suggesting that combining classification with segmentation and report generation supervision enhances the modeling of diffuse or subtle abnormality patterns. On `RAD-ChestCT`, UniCT achieves strong generalization, matching CT-Scroll [12] in F1-Score and AUROC while additionally providing segmentation masks and radiology reports.

Table 1: Abnormality classification performance on `ReXGroundingCT` [16], and (†)cross-dataset evaluation on `RAD-ChestCT` [13]. **Best results** are in bold, second-best are underlined. ■ Classification supervision. ♦ Multi-task.

	`ReXGroundingCT`			`RAD-ChestCT`(†)		
Model	F1-Score	AUROC	mAP	F1-Score	AUROC	mAP
■ ResNet3D [7]	.323 ±.006	.646 ±.006	.284 ±.008	.405 ±.008	.628 ±.028	.304 ±.004
■ ViViT [25]	.336 ±.010	.661 ±.011	.308 ±.017	.411 ±.008	.625 ±.012	.311 ±.002
■ CT-Net [13]	.371 ±.013	.698 ±.016	.336 ±.013	.416 ±.011	.644 ±.014	.304 ±.005
■ CT-Scroll [12]	.394 ±.011	.724 ±.007	.370 ±.055	**.427** ±.004	**.665** ±.004	.313 ±.003
♦ Multi-task base.	.389 ±.009	.731 ±.004	.365 ±.006	.422 ±.015	.654 ±.009	.314 ±.009
♦ UniCT (Ours)	**.410** ±.006	**.743** ±.003	**.383** ±.010	**.427** ±.007	**.665** ±.009	**.323** ±.009

3.2 Multi-Task Ablation Study

We further evaluate UniCT on the segmentation and report generation tasks, performing a cross-task ablation study to quantify the impact of MTF, SCF, and auxiliary task supervision. For comparison, we include *task-specific baselines* for each task: classification uses a ResNet 2.5D backbone applied to triplet-wise slices, followed by global average pooling before the classification head [21]; segmentation uses a U-Net [29]; report generation uses CT2Rep [17]. In all cases, the architectural components match those used in UniCT, isolating the effect of multi-task interactions. Table 2 reports F1-Score and AUROC for classification, foreground Dice (DSC) and Sensitivity for segmentation, and Clinical Accuracy RadBERT macro F1-Score (CA F1) [36] and CRG [18] for report generation.

Effect of joint multi-task learning. UniCT outperforms task-specific baselines across all metrics (paired t-test, $p < 0.01$), demonstrating that joint multi-task learning provides complementary supervision, and that shared spatial and semantic reasoning benefits all tasks. Although segmentation performance is impacted by the extreme sparsity of voxel-level annotations, as observed in `ReXGroundingCT` [2] and reflected in the ReXRank leaderboard [37], UniCT yields average Sensitivity gains of +4% for *non-focal* and +14% for *focal* classes over a U-Net, while achieving consistently higher DSC for bronchiectasis (.158, +17%), pleural effusion (.192, +24%) and ground-glass opacity (.213, +37%).

Effect of the fusion module (1). Removing MTF results in systematic performance drops across all metrics, suggesting that explicit latent interaction between task-specific representations is crucial for enhancing multi-task learning.

Effect of the segmentation modulation (2). SCF primarily benefits segmentation performance, as evidenced by decreases in DSC and Sensitivity when it is removed. Abnormality classification and automated report generation remain stable, indicating that SCF operates as a targeted mechanism that enhances dense prediction without perturbing global representations.

Effect of classification (3). Removing $\mathcal{L}_{\text{CLS}}$ degrades segmentation and report generation performance, suggesting that abnormality classification provides a strong global supervisory signal that stabilizes shared representations and facilitates effective learning for both dense spatial prediction and report generation.

Table 2: Comparative analysis of classification, segmentation and report generation performances on `ReXGroundingCT` [2]. **Best results** are in bold.

	Classification		Segmentation		Report generation	
Method	F1-Score	AUROC	DSC	Sensitivity	CA F1	CRG
Task-specific base.	.358 ±.011	.686 ±.016	.078 ±.007	.282 ±.008	.114 ±.016	.348 ±.005
Multi-task base.	.389 ±.009	.731 ±.004	.093 ±.007	.290 ±.010	.155 ±.013	.371 ±.009
UniCT (Ours)	**.410** ±.006	**.743** ±.003	**.113** ±.003	**.310** ±.006	**.175** ±.007	**.377** ±.003
(1) w/o MTF	.376 ±.013	.723 ±.008	.095 ±.006	.301 ±.011	.144 ±.017	.368 ±.005
(2) w/o SCF	.406 ±.009	.738 ±.010	.094 ±.011	.304 ±.005	.174 ±.009	.376 ±.004
(3) w/o $\mathcal{L}_{\text{CLS}}$	-	-	.094 ±.008	.293 ±.010	.152 ±.017	.364 ±.010
(4) w/o $\mathcal{L}_{\text{SEG}}$	.400 ±.004	.740 ±.006	-	-	.166 ±.016	.373 ±.002
(5) w/o $\mathcal{L}_{\text{REP}}$	.397 ±.012	.742 ±.003	.089 ±.007	.306 ±.005	-	-

Effect of segmentation (4). $\mathcal{L}_{\text{SEG}}$ improves abnormality classification and report generation performance. Despite the sparsity of voxel-level annotations in `ReXGroundingCT`, $\mathcal{L}_{\text{SEG}}$ encourages spatially-aware feature learning which benefits both discriminative classification and meaningful report generation.
Effect of report generation (5). Removing $\mathcal{L}_{\text{REP}}$ leads to decreased F1-Score for classification and DSC for segmentation, indicating that next-token prediction provides complementary semantic supervision, encouraging representations that are beneficial across both discriminative and dense prediction tasks.

3.3 Complementarity with Foundation Models

To disentangle representation scale from structured supervision, we replace our 2.5D encoder with frozen volumetric features ($H_{\text{CT}} \in \mathbb{R}^{864\times24\times24\times24}$) from the foundation model COLIPRI [35], adapting the segmentation modules to 3D and matching COLIPRI's preprocessing. We compare task-independent performance using these frozen features to our *UniCT with COLIPRI* framework. To avoid patient overlap, evaluation is restricted to the 172 `ReXGrounding` test subjects not included in the `CT-RATE` dataset used during COLIPRI pretraining. As shown in Table 3, UniCT improves performance across all tasks, with particularly large gains in clinical accuracy metrics for report generation, demonstrating that cross-task interaction provides complementary benefits beyond large-scale pretraining.

	Classification		Segmentation		Report generation	
Method	F1-Score	AUROC	DSC	Sensitivity	CA F1	CRG
COLIPRI (independent)	.417 ±.012	.771 ±.010	.075 ±.003	.136 ±.007	.138 ±.012	.351 ±.004
UniCT with COLIPRI	.427 ±.011	.791 ±.007	.096 ±.004	.158 ±.005	.206 ±.013	.383 ±.006

Table 3: Comparison between task-independent training and UniCT's unified multi-task interaction applied on identical frozen features from COLIPRI [35].

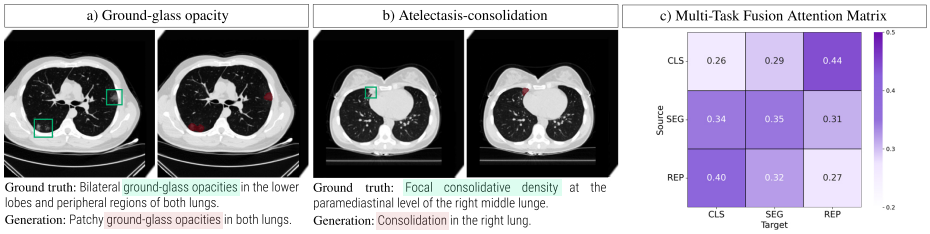

Fig. 2: (a & b): Examples of correctly classified abnormalities. Manually drawn reference boxes (green) are shown for visualization, with predicted segmentations (red). (c) MTF Attention Matrix averaged over heads, test samples and 5-folds.

3.4 Qualitative Results

Fig. 2 shows an example of correct predictions (a & b), illustrating that UniCT can jointly provide discriminative predictions alongside spatial grounding and clinically coherent textual descriptions. Analysis of the MTF attention matrix (c) reveals structured collaboration between tasks where classification and report generation exhibit strong bidirectional interactions, suggesting shared semantic reasoning, while segmentation distributes attention more uniformly across tasks, indicating its role as a spatial-semantic mediator within the unified framework.

4 Conclusion and Discussion

We propose UniCT, the first end-to-end unified framework that jointly addresses abnormality classification, segmentation, and report generation from 3D CT scans. By explicitly modeling cross-task interactions through multi-task supervision and feature fusion, UniCT produces complementary clinical outputs that align with radiologists' workflows, including abnormality prediction, spatial localization, and descriptive findings. Across two public datasets, UniCT matches or improves performance relative to comparable end-to-end methods with similar training regimes and capacity, suggesting that joint learning with explicit cross-task interaction constitutes a strong inductive bias for volumetric CT analysis.

Limitations & Future work. On the clinical side, despite evidenced mutual benefits of joint learning, the clinical performance metrics across all tasks remain below the acceptable thresholds in clinical practice. While UniCT is trained and evaluated on over 3,000 patients, broader datasets with richer spatial annotations would both improve performance and strengthen its generalization. Future work will explore scaling the framework to larger cohorts as additional annotations become available. On the technical side, although this work primary focuses on principled end-to-end joint learning, opportunities for improvement could come from integrating more expressive encoders and pretraining strategies within the same unified formulation. We anticipate that UniCT provides a flexible foundation for future extensions that incorporate more advanced grounding or reasoning mechanisms while preserving unified multi-task optimization.

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

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., Schmid, C.: ViViT: A Video Vision Transformer. In: ICCV. IEEE (2021)
2. Baharoon, M., Luo, L., Moritz, M., Kumar, A., Kim, S.E., Zhang, X., Zhu, M., Alabbad, M.H., Alhazmi, M.S., Mistry, N.P., Bijnens, L., Kleinschmidt, K.R., Rajpurkar, P., et al.: ReXGroundingCT: A 3D Chest CT Dataset for Segmentation of Findings from Free-Text Reports (Oct 2025), arXiv:2507.22030 [eess]
3. Baharoon, M., Ma, J., Fang, C., Toma, A., Wang, B.: Exploring the Design Space of 3D MLLMs for CT Report Generation. In: MICCAI (2025)
4. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Hyland, S.L., et al.: MAIRA-2: Grounded Radiology Report Generation (Sep 2024), arXiv:2406.04449 [cs]
5. Blankemeier, L., Cohen, J.P., Kumar, A., Veen, D.V., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., Chaudhari, A.S., et al.: Merlin: A Vision Language Foundation Model for 3D Computed Tomography (2024)
6. Broder, J., Warshauer, D.M.: Increasing utilization of computed tomography in the adult emergency department, 2000-2005. Emergency Radiology (Oct 2006)
7. Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset (Feb 2018), arXiv:1705.07750 [cs]
8. Chen, Z., Luo, L., Bie, Y., Chen, H.: Dia-LLaMA: Towards Large Language Model-driven CT Report Generation. In: MICCAI (2025)
9. Dai, P., Ou, Y., Yang, Y., Liu, D., Suzuki, K., et al.: SaSaMIM: Synthetic Anatomical Semantics-Aware Masked Image Modeling for Colon Tumor Segmentation in Non-contrast Abdominal Computed Tomography. In: MICCAI (2024)
10. DenOtter, T.D., Schubert, J.: Hounsfield Unit. In: StatPearls. StatPearls Publishing, Treasure Island (FL) (2024)
11. Di Piazza, T., Lazarus, C., Nempont, O., Boussel, L.: CT-AGRG: Automated abnormality-guided report generation from 3d chest ct volumes. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 01–05 (2025)
12. Di Piazza, T., Lazarus, C., Nempont, O., Boussel, L.: Imitating Radiological Scrolling: A Global-Local Attention Model for 3D Chest CT Volumes Multi-Label Anomaly Classification. In: MIDL (Jan 2025)
13. Draelos, R.L., Dov, D., Mazurowski, M.A., Lo, J.Y., Henao, R., Rubin, G.D., Carin, L.: Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. Medical Image Analysis p. 101857 (Jan 2021)
14. Du, Y., Bai, F., Huang, T., Zhao, B.: SegVol: Universal and Interactive Volumetric Medical Image Segmentation (Feb 2025), arXiv:2311.13385 [cs]
15. Gu, Y., Lei, W., Chen, H., Zhang, S., Zhang, X.: Interactive Segmentation and Report Generation for CT Images. In: MICCAI (2025)
16. Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasdelen, M.F., Durugol, O.F., Wittmann, B., Amiranashvili, T., Simsar, E., Simsar, M., Erdemir, E.B., Alanbay, A., Sekuboyina, A., Lafci, B., Bluethgen, C., Ozdemir, M.K., Menze, B.: Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography (Oct 2024), arXiv:2403.17834 [cs]

17. Hamamci, I.E., Er, S., Menze, B.: CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging (Mar 2024), arXiv:2403.06801 [cs, eess]
18. Hamamci, I.E., Er, S., Shit, S., Reynaud, H., Kainz, B., Menze, B.: CRG Score: A Distribution-Aware Clinical Metric for Radiology Report Generation (2025)
19. 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 (Oct 2025), arXiv:2510.20639 [cs]
20. Hao, J., Zhang, X., Liu, W., Yin, X., Gao, Y., Li, C., Zhang, L., Lu, L., Shi, Y., Han, X., Yan, K.: PLUS: Plug-and-Play Enhanced Liver Lesion Diagnosis Model on Non-Contrast CT Scans. In: MICCAI (2025)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (Dec 2015), arXiv:1512.03385 [cs]
22. Hosseini, A., Ibrahim, A., Serag, A.: From Slices to Volumes: Multi-Scale Fusion of 2D and 3D Features for CT Scan Report Generation. MICCAI (2025)
23. Li, F., Iyengar, A., Xu, L.: MTMed3D: A Multi-Task Transformer-Based Model for 3D Medical Imaging (2025), version Number: 1
24. Li, S., Qin, P., Wu, H., Nie, D., Thirunavukarasu, A.J., Yu, J., Zhang, L.: µ2 Tokenizer: Differentiable Multi-Scale Multi-Modal Tokenizer for Radiology Report Generation. In: MICCAI (2025)
25. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video Swin Transformer (Jun 2021), arXiv:2106.13230 [cs]
26. Pai, S., Hadzic, I., Bontempi, D., Bressem, K., Kann, B.H., Fedorov, A., Mak, R.H., Aerts, H.J.W.L.: Vision Foundation Models for Computed Tomography (2025)
27. Patel, P.R., De Jesus, O.: CT Scan. In: StatPearls. StatPearls Publishing, Treasure Island (FL) (2024)
28. Perez, E., Strub, F., Vries, H.d., Dumoulin, V., Courville, A.: FiLM: Visual Reasoning with a General Conditioning Layer (Dec 2017), arXiv:1709.07871 [cs]
29. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation (May 2015), arXiv:1505.04597 [cs]
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge (Jan 2015), arXiv:1409.0575 [cs]
31. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations (2017)
32. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and Explainable Region-guided Radiology Report Generation. In: CVPR (2023)
33. Tian, Y., Song, Y.: Feature Decomposition via Shared Low-rank Matrix Recovery for CT Report Generation. IEEE Transactions on Medical Imaging (2025)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (Aug 2023), arXiv:1706.03762 [cs]
35. Wald, T., Hamamci, I.E., Gao, Y., Bond-Taylor, S., Sharma, H., Pérez-García, F., et al.: Comprehensive language-image pre-training for 3D medical image understanding (Jan 2026), arXiv:2510.15042 [cs]
36. Yan, A., McAuley, J., Lu, X., Du, J., Chang, E.Y., Gentili, A., Hsu, C.N.: RadBERT: Adapting Transformer-based Language Models to Radiology. Radiology: Artificial Intelligence (4), e210258 (Jul 2022)
37. Zhang, X., Zhou, H.Y., Yang, X., Banerjee, O., Acosta, J.N., Miller, J., Huang, O., Rajpurkar, P.: ReXrank: A Public Leaderboard for AI-Powered Radiology Report Generation (Nov 2024), arXiv:2411.15122 [cs]

38. Zhang, Y., Li, H., Du, J., Qin, J., Wang, T., Chen, Y., Lei, B.: 3D Multi-Attention Guided Multi-Task Learning Network for Automatic Gastric Tumor Segmentation and Lymph Node Classification. IEEE Transactions on Medical Imaging (2021)
39. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhou, X., Zhang, Y., Wang, Y., Xie, W.: Large-vocabulary segmentation for medical images with text prompts. npj Digital Medicine (1), 566 (Sep 2025)