RadGPT: Constructing 3D Image-Text Tumor Datasets

Pedro R. A. S. Bassi[1,2,3] Mehmet Can Yavuz[4] Ibrahim Ethem Hamamci [6,7] Sezgin Er [5]
Xiaoxi Chen[8] Wenxuan Li[1] Bjoern Menze[6] Sergio Decherchi[3]
Andrea Cavalli[2,3,9] Kang Wang[4] Yang Yang[4] Alan Yuille[1] Zongwei Zhou[1,*]

[1]Johns Hopkins University [2]University of Bologna [3]Italian Institute of Technology
[4]University of California, San Francisco [5]Istanbul Medipol University [6]University of Zurich
[7]ETH AI Center [8]University of Illinois Urbana-Champaign
[9]École Polytechnique Fédérale de Lausanne

Code, dataset, and models: https://github.com/MrGiovanni/RadGPT

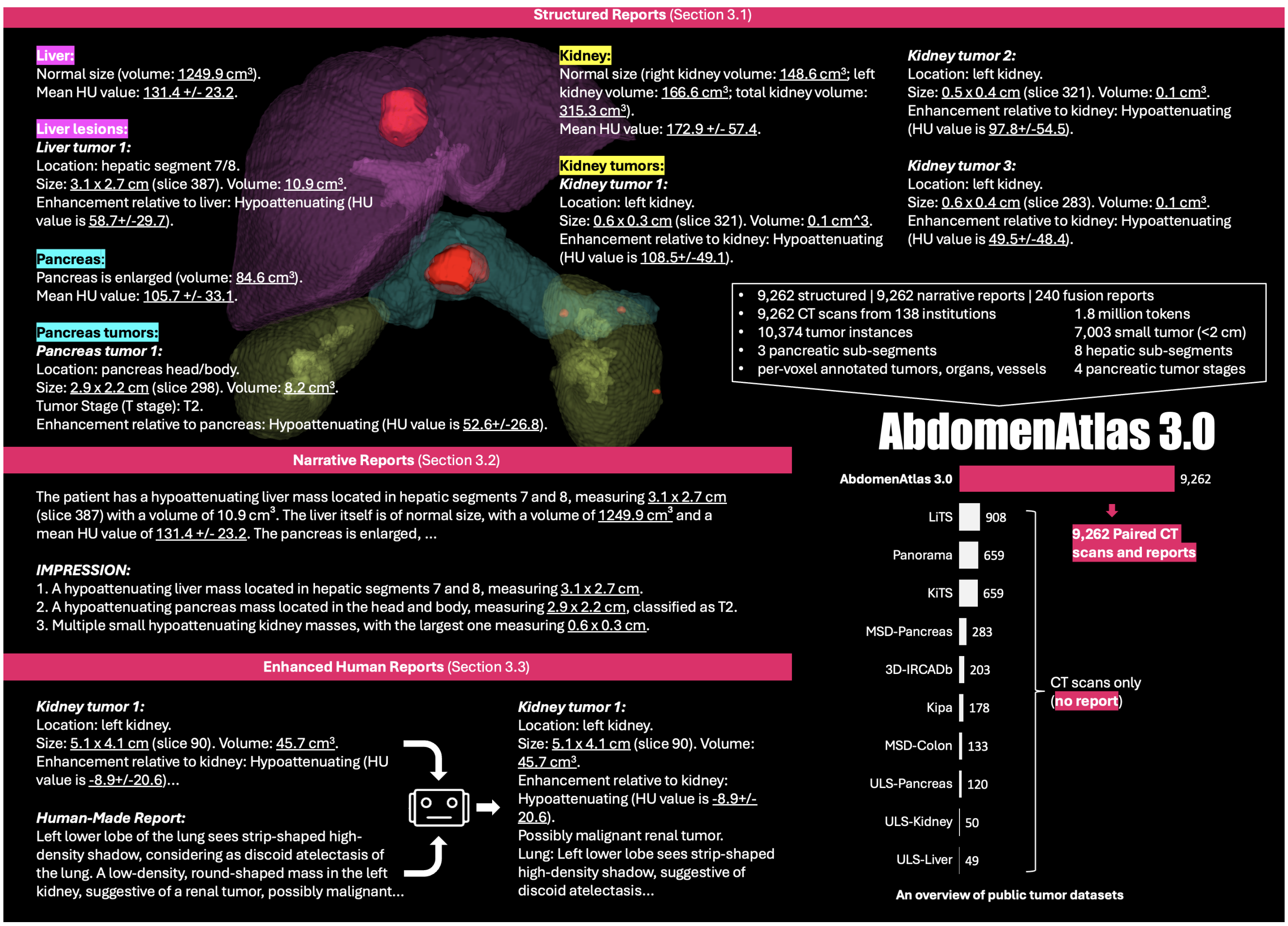

Figure 1. **AbdomenAtlas 3.0 is a large-scale, image-text tumor dataset of 9,262 3D CT scans.** Each CT scan has per-voxel tumor annotations and reports, including 5,582 liver tumors, 368 pancreatic tumors and 4,424 kidney tumor, 7,003 of which are small tumors (≤2cm). In addition, AbdomenAtlas 3.0 provides detailed annotations for pancreatic cancer staging (T1–T4), as well as per-voxel segmentation of liver sub-segments (1–8) and pancreatic sub-segments (head, body, and tail). Structured, narrative, and enhanced reports were created by a team of 12 board-certified radiologists assisted by our proposed Radiology Generative Pretrained Transformer (RadGPT).

Abstract

*Cancers identified in CT scans are usually accompanied by detailed radiology reports, but publicly available CT datasets often lack these essential reports. This absence limits their usefulness for developing accurate report generation AI. To address this gap, we present **AbdomenAtlas 3.0**, the first public, high-quality abdominal CT dataset with detailed, expert-reviewed radiology reports. All reports are paired with per-voxel masks and they describe liver, kidney and pancreatic tumors. AbdomenAtlas 3.0 has 9,262 triplets of CT, mask and report—3,955 with tumors. These CT scans come from 17 public datasets. Besides creating the reports for these datasets, we expanded their number of tumor masks by 4.2×, identifying 3,011 new tumor cases. Notably, the reports in AbdomenAtlas 3.0 are more standardized, and generated faster than traditional human-made reports. They provide details like tumor size, location, attenuation and surgical resectability. These reports were created by **12** board-certified radiologists using our proposed **RadGPT**, a novel framework that converted radiologist-revised tumor segmentation masks into structured and narrative reports. Besides being a dataset creation tool, RadGPT can also become a fully-automatic, segmentation-assisted report generation method. We benchmarked this method and 5 state-of-the-art report generation vision-language models. Our results show that segmentation strongly improves tumor detection in AI-made reports.*

1. Introduction

Each year, over 85 million CT scans are performed in the United States [44, 52], growing 6% per year, and significantly outpacing the 0.7% annual growth rate of the medical imaging workforce [13]. This disparity puts radiologists under significant time pressure, making it challenging to generate detailed, accurate radiology reports. AI may support report generation, but it requires data. To address this gap, we present AbdomenAtlas 3.0 (summarized in Figure 1 and Table 1), the first high-quality abdominal CT dataset with reports. It has 9,262 3D CTs in NIfTI format (2,789,975 CT slices) sourced from 17 public datasets (Table 1), which originally had no radiology report. 12 board-certified radiologists, assisted by RadGPT (introduced below), generated reports for all CTs—totaling 1,843,262 tokens. For each CT, we document tumor size, location, attenuation (HU), and volume for each identified tumor. Reports also include T-stage for pancreatic cancer (PDAC), derived from tumor size and vessel involvement, critical for surgery. Each CT has both structured (template-based) and narrative (free-text) reports, and precise voxel-level annotations. Reports cover tumors in the liver, pancreas, and kidneys, including 3,011 tumors newly identified by the radiologists. Our reports also describe organ abnormalities (e.g., fatty liver, enlarged spleen), patient demographics, and contrast phase. They locate tumors in liver segments (1–8) and pancreas segments (head, body, tail)—all annotated per-voxel. This is the largest liver sub-segment dataset, and the first public pancreas sub-segment dataset. Also, we enhanced 240 existing human-made reports, covering 66 distinct diagnoses, with more detailed tumor analyses.

To create AbdomenAtlas 3.0, we developed Radiology Generative Pre-trained Transformers (**RadGPT**), an anatomy-aware vision-language AI agent that assists radiologists in creating CT-report datasets. We started with our previous **AbdomenAtlas 1.1** [31], composed of 17 public datasets and their organ segmentation masks, but no tumor segmentation nor report. First, RadGPT segments liver, kidney, and pancreas tumors, along with liver/pancreas sub-segments, surrounding organs, and blood vessels[1]. Then, radiologists revise the segmented tumors, annotating missed ones and removing false positives. We call the dataset with CT scans and tumor segmentation masks **AbdomenAtlas 2.0**, and we also release it here. From the revised segmentations, RadGPT extracts attributes (e.g., tumor size, volume, attenuation, stage) via deterministic, rule-based algorithms. These attributes are used to fill a radiologist-designed template, producing *structured reports*. RadGPT's deterministic algorithms ensure that the structured reports are fully explainable and fully coherent with the radiologist-revised segmentations. Next, RadGPT converts the structured reports into free-text *narrative reports*, using large language models (LLMs) that emulate the style (word choice and organization) of radiologists at a major US hospital—through in-context learning with special example selection (§3.2). Last, RadGPT fuses per-voxel segmentations with human-made reports/clinical notes to produce *enhanced human reports* (§3.3), combining precise and detailed tumor analysis from segmentation with broader diagnostic range (66 diagnoses) from human-made reports. Reports were verified by radiologists (Appendix C). We call the final triplet dataset—CT scans, tumor masks, reports—**AbdomenAtlas 3.0**.

We evaluated six CT report generation models on AbdomenAtlas 3.0 (internal validation) and a private dataset (external validation): CT2Rep [21], M3D [4], CT-CHAT [20], Merlin [8], RadFM [54] and RadGPT. Besides a dataset creation tool, RadGPT can also become a fully-automatic, segmentation-assisted report generation model, by converting the outputs of a segmentation model into reports, without radiologist revision. We expect AbdomenAtlas 3.0 to foster segmentation-assisted report generation, as the dataset has CTs, per-voxel annotations and reports. We evaluated all report generation models with a new diagnostic metric (§B.3). It first uses an LLM to extract labels (tumor presence) from AI- and human-made reports. Then, it compares the labels from AI- and human-made re-

dataset	CTs	institutions	countries	annotated liver tumors	annotated pancreatic tumors	annotated kidney tumors
FLARE'23 [2022] [link]	4,100	35	1	0 → 564	0 → 38	0 → 941
KiTS'23 [2020] [link]	489	1	1	0 → 1	0	452
LiTS [2019] [link]	131	7	5	50	0	0
TCIA-Pancreas-CT [2015] [link]	42	1	1	0	0	0
CT-ORG [2020] [link]	140	8	6	0 → 44	0	0 → 21
Trauma Det. [2023] [link]	4,714	23	13	0 → 113	0 → 32	0 → 38
BTCV [2015] [link]	47	1	1	0	0	0
CHAOS [2018] [link]	20	1	1	0	0 → 1	0
AbdomenCT-1K [2021] [link]	1,050	12	7	0 → 117	0 → 94	0 → 181
MSD CT Tasks (6) [2021] [link]	945	1	1	251 → 462	191	0 → 388
WORD [2021] [link]	120	1	1	0 → 47	0 → 1	0 → 45
AMOS [2022] [link]	200	2	1	0 → 74	0 → 4	0 → 56
AbdomenAtlas 3.0 (ours)	**9,262**	**138**	**19**	**301 → 1,472**	**191 → 361**	**452 → 2,122**

dataset	liver sub-segments	pancreas sub-segments	peripancreatic blood vessels[1]	tumor stage	radiology reports	text tokens
FLARE'23 [2022] [link]	✗	✗	✗	✗	0	0
KiTS'23 [2020] [link]	✗	✗	✗	✗	0	0
LiTS [2019] [link]	✔	✗	✗	✗	0	0
TCIA-Pancreas-CT [2015] [link]	✗	✗	✗	✗	0	0
CT-ORG [2020] [link]	✗	✗	✗	✗	0	0
Trauma Det. [2023] [link]	✗	✗	✗	✗	0	0
BTCV [2015] [link]	✗	✗	✗	✗	0	0
CHAOS [2018] [link]	✗	✗	✗	✗	0	0
AbdomenCT-1K [2021] [link]	✗	✗	✗	✗	0	0
MSD CT Tasks (6) [2021] [link]	✗	✗	✗	✗	0	0
WORD [2021] [link]	✗	✗	✗	✗	0	0
AMOS [2022] [link]	✗	✗	✗	✗	0	0
AbdomenAtlas 3.0 (ours)	✔	✔	✔	✔	**18,524**	**1,843,262**

→ represents the number of CT scans with tumor annotations in the original dataset, followed (→) by our updated number of CT scans with tumor annotations, including the additional annotations AbdomenAtlas 3.0 provided with radiologist support.

Table 1. **Besides being the only public abdominal CT dataset with paired radiology reports, AbdomenAtlas 3.0 offers 4.2× more annotated tumors than the combined total of its constituent datasets**. The table highlights how AbdomenAtlas 3.0 enhances public datasets with reports and tumor annotations. It includes 1,472 CT scans with liver tumors, 361 with pancreatic tumors, and 2,122 with kidney tumors, most newly annotated with radiologist support. Each sample includes per-voxel annotations and reports. AbdomenAtlas 3.0 is also the first dataset to provide per-voxel segmentations of pancreas sub-segments and peripancreatic blood vessels. AbdomenAtlas 1.1 [31] has the same CTs as AbdomenAtlas 3.0, but it has only organ segmentation masks—no tumor masks, reports, organ sub-segments, nor peripancreatic blood vessels. AbdomenAtlas 2.0 has the same CTs and masks as 3.0, no report—it is our intermediate step before 3.0.

ports to evaluate AI's diagnostic sensitivity and specificity (§3.4). To validate this new metric, radiologists manually evaluated LLM labeling—it achieved 96% zero-shot accuracy (Figure 4). Our contributions are:

1. AbdomenAtlas 3.0 is the first public dataset with high-quality abdominal CT scans (9,262), radiology reports (structured, narrative, and enhanced), and tumor masks.
2. With 12 radiologists, we annotated 3,011 new tumors in the 17 public datasets inside AbdomenAtlas 3.0—expanding their number of tumor masks by 4.2×.
3. Our reports locate liver and pancreas tumors within sub-segments of the organs. They also measure contact between tumors and blood vessels for pancreatic tumor staging. Staging and sub-segments are key for surgery.
4. We developed Rad-GPT to assist dataset creation: unlike current VLMs, it uses deterministic algorithms to convert radiologist-revised tumor masks into reports, improving reports' trustworthiness and interpretability. Also, RadGPT can generate fully-automated reports.
5. We benchmarked 5 SOTA VLMs for report generation and showed segmentation improves report generation.

2. Related Work

Per-voxel tumor annotations are scarce. Most public abdominal CT datasets concentrate on a single tumor type (e.g., liver [7], pancreas [3], or kidney [22]) and contain only a few hundred tumor annotations (Table 1). This small volume of annotations hinders effective AI training and evaluation. To address this, our radiologists have quadrupled the number of per-voxel tumor annotations in the 17 public datasets included in AbdomenAtlas 3.0 (Table 1).

Real-world radiology reports are even rarer than per-voxel tumor annotations. At the time of writing, no publicly available abdominal CT dataset contains authentic clinical reports. Only one dataset, M3D-Cap [4], provides textual captions (sourced from Radiopaedia [19]), but its scans are 2D JPG/PNG image series rather than standard 3D NIfTI or DICOM volumes. Consequently, crucial information such as inter-slice spacing and Hounsfield units (HU) is missing [59]. In contrast, CT scans in AbdomenAtlas 3.0 were collected in standard formats from 138 medical institutions, retaining clinically important metadata. As an-

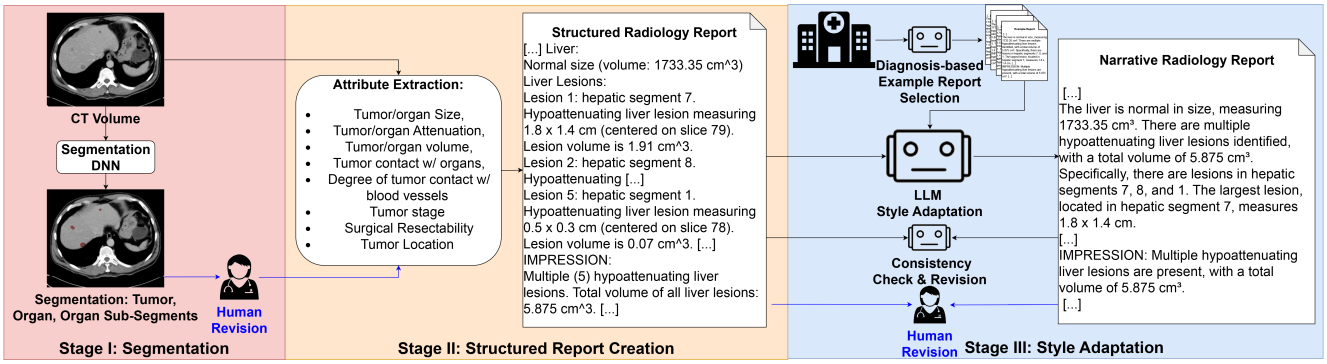

Figure 2. **The RadGPT 3-stage pipeline for report generation.** Blue arrows denote human revision used to create AbdomenAtlas 3.0. By skipping revision, RadGPT can also become a fully-automatic, segmentation-assisted report generation AI. **Stage I. Segmentation.** DiffTumor [11] and nnU-Net [25] segment 26 anatomical structures important for cancer detection and staging[1]. Radiologists corrected wrong tumor segmentations in AbdomenAtlas 3.0, and ground-truths from public datasets were used when available. **Stage II. Structured Report Generation.** Deterministic algorithms (§3.1.1-3.1.3) extract radiologist-selected attributes—important for cancer detection, staging and treatment—from CTs and segmentations. Attributes fill a radiologist-designed template, generating structured reports detailing liver, kidney, and pancreatic tumors. The rule-based deterministic algorithms ensure the reports are fully coherent with segmentations and explainable. **Stage III. Style Adaptation.** LLM adapts structured reports into a target hospital's narrative style, leveraging example reports from the hospital—in-context learning prioritizing examples of similar diagnoses (§3.2). LLM is asked to preserve medical information and double checks for consistency. Radiologists revised reports in AbdomenAtlas 3.0. Also, LLM can fuse structured and human-made reports, creating enhanced human reports combining segmentation-based precision with humans' broad diagnostic range (§3.3).

other unique quality, our reports are paired with tumor and organ masks—fostering the development of segmentation-assisted report generation AI.

Due to the scarcity of reports in public datasets, only two models specifically target abdominal CT report generation: M3D [4] (publicly released) and Merlin [8] (partially released). Text-similarity metrics were used to evaluate both models (e.g., BLEU and ROUGE [35]; Merlin was also evaluated with RadGraph-F1), but these metrics can be skewed by style variations even when the underlying diagnoses remain unchanged (§B.3). In contrast, we propose the evaluation of AI-generated reports using diagnostic sensitivity and specificity (Table 2)—clinically meaningful and acceptable metrics [9, 55]. Lastly, although many report-generation models exist for 2D X-ray [12, 33, 34, 45, 51, 57], adapting them to 3D CT may require profound re-design, which may unfairly represent the originals. *Why?* ***First***, tumors in CTs can occupy $\leq$0.0001% of the full volume, vs. 5–10% in X-rays. ***Second***, many X-ray models rely on 2D pre-trained models, but CT data is 3D. Processing CT slices individually is computationally prohibitive, and it is difficult to align slices with findings in reports. Thus, all models we evaluated in AbdomenAtlas 3.0 [4, 8, 20, 21, 54] are designed for CT.

3. AbdomenAtlas 3.0 & RadGPT

Table 1 shows advantages of **AbdomenAtlas 3.0** over its 17 source datasets—*providing reports, organ sub-segments and blood vessels annotated per-voxel, and 4$\times$ more tumor annotations*. Sections §3.1–§3.3 explain RadGPT (summarized in Figure 2), and how it empowered 12 radiologists to generate reports for the 9,262 CTs in AbdomenAtlas 3.0.

3.1. Creating Structured Reports

Structured reports use a radiologist-designed template, enhancing clarity and aiding medical decisions [1] (Figure 1). To fill the template, RadGPT uses segmentation and deterministic algorithms to: (1) sub-segment organs to locate tumors (§3.1.1); (2) measure tumor size, volume, and attenuation (§3.1.2); (3) perform cancer staging from tumor and blood vessel segmentations (§3.1.3).

3.1.1. Sub-segment Organs to Locate Tumors

Human-made reports use organ sub-segments to locate tumors. Location is key for prognosis, tracking tumor progression, and treatment planning. E.g., the possibility of tumor surgical removal depends on its location [47]. To locate liver and pancreas tumors in structured reports, RadGPT sub-segments the organs and checks which sub-segments intersect with the tumor. RadGPT segments tumors with DiffTumor [11], a public segmentation model, and radiologists revise the segmentations (Appendix C).

For liver sub-segmentation, we leverage whole-liver ground-truth per-voxel annotations to help the AI find liver sub-segments. First, we offset the liver intensity (by 200 HU), following its ground-truth per-voxel annotation. Second, using these CT scans with offsets as input, we trained an nnU-Net [25] for liver sub-segmentation. The sub-segments follow the Couinaud standard [15], which divides the liver into eight sub-segments that can be independently removed in surgeries. Couinaud sub-segment annotations are publicly available for 131 LiTS CT scans [7, 58], which we used for training. Given the small size of this dataset, we fine-tuned an nnU-Net pre-trained on 9,262 CT scans in AbdomenAtlas 1.1 [31]. After fine-tuning, we inferenced the nnU-Net on AbdomenAtlas 3.0. The HU value offsetting ensured the precise alignment between the generated sub-segments and the existing liver ground-truth per-voxel annotations. AbdomenAtlas 3.0 is the second [58] but *largest public dataset with liver sub-segments*.

For pancreas sub-segmentation, there is no public dataset with per-voxel annotations of pancreas head, body, and tail—ours is the first. Thus, to subsegment the pancreas, we used the superior mesenteric artery (SMA) as a landmark. We trained an nnU-Net to segment the SMA (using private data) and developed a deterministic algorithm that uses the SMA segmentation to sub-segment the pancreas (Sup. Alg. 1). First, it uses the SMA to find the pancreatic neck, since it curves around the SMA. The neck locates the head-body boundary. Then, the body-tail boundary is set at the midpoint along their length. Our landmark-based deterministic algorithm closely mimics how radiologists use mesenteric vessels to subsegment the pancreas [48]. AbdomenAtlas 3.0 is the *first public dataset with pancreas sub-segments*.

3.1.2. Measure Tumors Like Radiologists

Radiologists commonly measure tumors using use the World Health Organization (WHO) standard, which provides two diameters: the longest tumor diameter in any axial plane (D), and its perpendicular diameter in the same plane (d) [40]. Standardization of measurements is key for accurate cancer prognosis and treatment planning [32, 40]. Thus, RadGPT also uses the WHO standard, measuring tumors like radiologists. AbdomenAtlas 3.0 presents radiologist-revised segmentations of liver, kidney and pancreas tumors. From segmentations, RadGPT extracts tumor measurements using a deterministic algorithm that implements the WHO standard (Sup. Alg. 2). Besides diameters, our structured reports present tumor & organ *volume* and *attenuation* (HU values), also extracted from segmentation[1]. Using volumes, our reports diagnose enlarged organs, and attenuation diagnoses fatty liver (average HU < 40 [28]) and pancreas (pancreas-to-spleen attenuation < 0.7 [18])—a condition related to diabetes and pancreatic cancer [18]. Meanwhile, tumor attenuation helps identify tumor type.

[1] AbdomenAtlas 3.0 is the first dataset with per-voxel annotations for the blood vessels key for pancreatic tumor staging: the celiac axis (CA), superior mesenteric artery (SMA), superior mesenteric vein (SMV), common hepatic artery (CHA), and portal vein. These annotations were produced by an nnU-Net trained in private data, and revised by radiologists (Appendix C). AbdomenAtlas 3.0 also has per-voxel annotations for other 22 structures important for cancer detection/staging: liver tumors, kidney tumors, pancreas tumors, liver, kidney, pancreas, spleen, adrenal glands, stomach, duodenum, bile duct, intestines, aorta, and postcava.

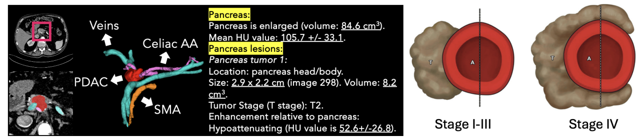

Figure 3. **Automated T staging.** Our RadGPT first segments the tumor and key vascular structures from CT scans, then measures tumor size and blood vessel contact angle to automatically assign T stage and resectability. If the tumor-vessel contact angle surpasses 180 degrees, the tumor becomes unresectable (T stage 4).

3.1.3. Stage Pancreatic Cancer using Segmentation

Tumor T-stage summarizes tumor size and relationship to nearby structures. It is key for surgical planning and survival, especially for pancreatic adenocarcinoma (PDAC), an aggressive cancer [1]. However, staging is time-consuming. As shown in Figure 3, for PDAC staging, radiologists must measure tumors (§3.1.2) and analyze its interaction with blood vessels (SMA, CHA, CA, SA) [1]. Accordingly, RadGPT first segments vessels and tumors (using nnU-Net and DiffTumor[1]) and radiologists revise segmentations (Appendix C). Then, a deterministic algorithm uses the revised segmentations to measure the tumor-vessel contact angle (Sup. Alg. 3). Large angles ($>180°$) make surgery difficult, increasing stage. For interpretability, reports justify stages with tumor size and tumor-vessel degree of contact, and our deterministic algorithm faithfully implements the guidelines radiologists use to stage PDAC [1]. AbdomenAtlas 3.0 is *first public dataset with PDAC T stage labels*.

3.2. Creating Narrative Reports

Structured reports use rigid templates to improve clarity and clinical decision-making [1]. However, rigid templates may conflict with the reporting style of an institution. Thus, RadGPT can create narrative reports that mimic the style of a target institution. In AbdomenAtlas 3.0, they mimic human-made reports at a major US hospital (Figure 1). The narrative reports are created through style adaptation with in-context learning: we provide a pre-trained LLM (Llama-3.1 70B, AWQ quantization [17]) with a structured report and 10 human-made reports from the target institution, and the LLM adapts the structured report to the style of the human-made reports. We ask the LLM *not* to change diagnoses or details. Thus, narrative reports contain all the detailed information from structured reports (§3.1).

However, style of human-made reports varies with diagnoses. E.g., pancreatic tumor differ from liver tumor reports [41]. Thus, we verify diagnoses to give the LLM example reports with the correct style. First, another LLM categorizes human-made reports according to tumors (liver, pancreas, kidney, none). Then, when adapting a structured report to narrative, the first LLM receives example human-made reports with the same tumor as the structured report.

Internal validation on the test set of AbdomenAtlas 3.0 (IID)

	pancreatic tumor (%)			kidney tumor (%)			liver tumor (%)		
Model	Sen. ($\leq$2 cm)	Sen. (>2 cm)	Spec.	Sen. ($\leq$2 cm)	Sen. (>2 cm)	Spec.	Sen. ($\leq$2 cm)	Sen. (>2 cm)	Spec.
CT-CHAT [20]	**66.7**	51.9	61.2	31.1	32.8	74.2	5.7	3.2	94.7
CT2Rep [21]	0.0	0.0	92.5	36.5	39.3	70.4	35.8	49.2	70.4
M3D [4]	0.0	7.4	97.2	8.1	16.4	84.1	9.4	12.7	86.0
Merlin [8]	33.3	51.9	71.8	28.4	45.9	86.6	30.2	41.3	**95.9**
RadFM [54]	0.0	0.0	**99.9**	3.7	6.3	**95.6**	3.3	5.7	93.9
RadGPT (ours)	**66.7**	**81.5**	93.2	**54.8**	**93.3**	51.8	**39.6**	**96.8**	64.4

External validation on unseen hospital—UCSF (OOD)

Model	Sen. ($\leq$2 cm)	Sen. (>2 cm)	Spec.	Sen. ($\leq$2 cm)	Sen. (>2 cm)	Spec.	Sen. ($\leq$2 cm)	Sen. (>2 cm)	Spec.
CT-CHAT [20]	27.5	N/A	73.1	24.3	29.7	74.6	5.2	4.2	94.0
CT2Rep [21]	2.1	N/A	96.7	4.0	10.0	98.0	0.0	0.0	**100.0**
M3D [4]	3.3	N/A	97.9	14.8	13.1	86.3	10.7	17.3	87.3
Merlin [8]	7.5	N/A	**100.0**	8.1	9.2	**100.0**	9.1	19.2	**100.0**
RadFM [54]	0.0	N/A	**100.0**	7.5	6.8	90.9	10.9	11.1	85.0
RadGPT (ours)	**76.9**	N/A	76.6	**92.0**	**97.3**	78.3	**79.6**	**89.4**	73.4

Table 2. **In tumor detection, fully-automated reports by RadGPT surpass reports created by end-to-end report generation models.** We use RadGPT as a fully-automated segmentation-assisted report generation model (Figure 2). The results indicate that per-voxel segmentation (step 1 in the RadGPT pipeline) may strongly improve report generation. We tested out-of-distribution (OOD) at UCSF, a hospital not seen in training, and in-distribution (IID). In the IID set, the ground-truth contained 9 small and 27 large pancreatic tumors, 74 small and 61 large kidney tumors, and 53 small and 63 large liver tumors, with 890, 791, and 810 negatives, respectively. In OOD, we have 385 (small) and 0 (large) for pancreas, 50 (small) and 219 (large) for kidney, and 142 (small) and 301 (large) for liver, with 244 negatives for each organ. Decision thresholds are analyzed in 15. While other methods were evaluated zero-shot, CT-CHAT, CT2Rep and Merlin were trained in AbdomenAtlas 3.0, giving them an advantage in the IID dataset. To compute sensitivity and specificity, we used our proposed diagnostic evaluation (§3.4): an LLM extracted binary tumor presence labels *per-organ*, and we compared the labels for AI-made reports and ground-truth human-made reports. LLM label extraction accuracy is 96% (Figure 4). Table 5 provides additional metrics (BLEU, ROUGE, BERT, RadGraph-F1), showing they are usually sensible to variations in report style, unlike our diagnostic evaluation.

After adapting a structured report into a narrative report, the LLM performed a quality check. It extracted diagnoses and quantitative information (e.g., tumor size and stage) from both reports and checked for consistency. We prompted the LLM to correct in the narrative report any information diverging from the structured report, and to remove any diagnosis not present in the structured report.

3.3. Creating Enhanced Human Reports

Like most abdominal CT datasets, AbdomenAtlas 3.0 focuses on tumors—as cancer is a major cause of death. Our reports can precisely measure and analyze multiple tumors in a CT, while human-made reports usually measure the largest tumors only (§4.3). However, human-made reports cover multiple diagnoses unrelated to tumors. To combine their strengths, RadGPT prompts the zero-shot LLM (Llama 3.1 70B AWQ) to fuse the details in structured reports with the many diagnoses in human-made reports/clinical notes (Figures 12 and 1), generating *enhanced human reports*. AbdomenAtlas 3.0 has 240 of them: 209 used clinical notes for TotalSegmentator CT scans [4], and 31 used notes from our radiologists. They span 66 diagnoses.

3.4. Evaluating Diagnoses in AI-Made Reports

We propose a new strategy to evaluate the clinical utility of AI-made reports: a straightforward, LLM-based diagnostic evaluation. First, we prompt a zero-shot LLM (Llama 3.1 70B AWQ, prompts in §B.4) to identify in which organ the report mentions tumors. Then, we convert the LLM answer into categorical labels. We compare labels for AI-made and human-made reports (ground-truth) to calculate tumor detection sensitivity and specificity. This evaluation strategy is *scalable* and practical: with zero-shot inference, it does not need fine-tuning and is easily adaptable to multiple hospitals. Importantly, our strategy produces clinically relevant metrics (detection sensitivity / specificity), which are easy to interpret by clinicians. Here, we limit our evaluation strategy to tumor detection. However, it can be expanded to evaluate other relevant clinical information and diseases beyond tumors—with simple prompt modifications.

4. Experiment & Result

We randomly selected 10% of AbdomenAtlas 3.0 as a test set, where we evaluated 6 CT report generation models: CT2Rep [21], M3D [4], CT-CHAT [20], Merlin [8], RadFM [54] and RadGPT—as baselines for future work. In AbdomenAtlas 3.0, RadGPT transforms radiologist-revised tumor masks into reports. In this section, we evaluate RadGPT as a fully-automated, segmentation-assisted method, without radiologist revision (Figure 2). We have both AI trained on AbdomenAtlas 3.0 (CT2Rep and CT-CHAT, see Appendix B.1 for training details) and those

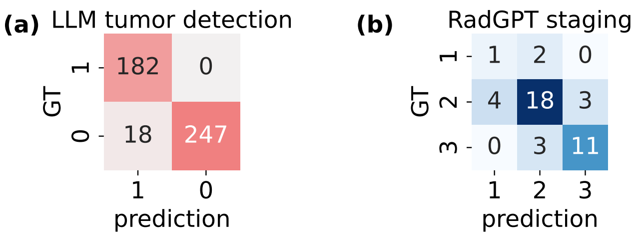

Figure 4. **Confusion matrices. (a)** A zero-shot LLM (Llama 3.1 70B AWQ) has 96% accuracy, 0.953 F1-score in determining if radiology reports show tumors. Thus, the LLM can accurately calculate tumor detection sensitivity and specificity for AI-made reports (§3.4). LLM's accuracy rivals established labelers, as those in CheXpert [24] and CheX-ray14 [50]. Results were manually evaluated by radiologists on 447 reports with kidney, pancreas, and liver tumors. **(b)** PDAC staging confusion matrix for RadGPT, the first public AI for staging abdominal CT tumors. Results on a private dataset with ground-truth tumor stage annotations (N=42).

trained on abdominal CTs in other works (M3D, Merlin, DiffTumor inside RadGPT). To ensure realistic evaluation [5, 6], we evaluate on the AbdomenAtlas 3.0 test set and on a private out-of-distribution (OOD) dataset, from the University of California San Francisco hospital (UCSF, California, USA) never seen by any AI in training.

Zero-shot LLMs can accurately evaluate report generation. For automated evaluation on a large test dataset, we will use an LLM (Llama-3.1) to assess the reports generated by 6 AI models (§3.4). Before that, radiologists verified the LLM's ability to determine whether a report indicates tumors or not. They *read the zero-shot LLM answers for 447 different reports*, verifying that it achieved 96% accuracy (Figure 4). Results demonstrate the LLM reliability in evaluating tumor detection, per-organ.

Segmentation can assist report generation models. The LLM-based evaluation, (Table 2) showed that the reports generated by RadGPT strongly surpassed the other abdominal CT report generation models, especially in the OOD test set (unseen hospital)[2]. End-to-end trained methods had difficulty detecting tumors in the OOD dataset (low sensitivity), and RadGPT strongly outperformed them for small and large tumors in the liver, pancreas, and kidneys. This performance difference shows the benefits of using segmentation to improve report generation: DiffTumor produces accurate tumor segmentations, which RadGPT translates into reports. By releasing AbdomenAtlas 3.0, the first abdominal CT dataset with triplets of CT scans, reports, and per-voxel annotations, our objective is to catalyze further research on segmentation-assisted report generation.

RadGPT is the first public AI model to perform cancer staging on abdominal CT. Figure 4 shows the performance of RadGPT for staging of pancreatic adenocarcinoma. RadGPT fully-automated reports achieved accuracy of 71.43% in determining tumor T stages 1 to 3. The results show that AI is a promising tool for assisting cancer staging, a key but time-consuming task for radiologists. Still, these fully-automatic results show radiologist revision is essential to ensure staging accuracy in AbdomenAtlas 3.0.

[2]As RadGPT narrative and structured reports match in diagnostic accuracy we present only one result for RadGPT.

	liver tumor	pancreatic tumor	kidney tumor
Detection Precision (%)	92.3 (12/13)	50.0 (8/16)	91.7 (11/12)
Size Accuracy (%)	100.0 (12/12)	75.0 (6/8)	100.0 (11/11)

Table 3. **RadGPT has 75.6% tumor detection precision and 93.5% tumor measurement accuracy.** A radiologist manually evaluated reports RadGPT created for 23 external test CTs (UCSF). A tumor measurement was considered correct if it deviated by $\leq$10% from the radiologist's measurement (both use the WHO measuring standard [40]). As evaluation is time-consuming, the radiologist evaluated 23 reports. Using an LLM for automatically evaluating tumor measurements is challenging: it requires pairing tumors in AI-made reports and ground-truth reports.

4.1. RadGPT Accurately Measures Tumor Size

An expert radiologist manually evaluated structured reports generated by RadGPT. He analyzed each reported tumor, evaluating its measurement and checking if the tumor is a false-positive (tumor not present in the CT volume) or a true-positive (present). The radiologist deemed 75.6% of the tumors reported by RadGPT true-positives, and 93.5% of them were correctly measured (Table 3). RadGPT only made measuring mistakes for pancreatic tumors (PDAC), but even the radiologist could not measure 3 PDACs.

4.2. RadGPT Locates Tumors in Organs

RadGPT uses use organ sub-segments to locate tumors. It achieved a Dice similarity coefficient (DSC) of 0.85 in segmenting eight liver sub-segments, according to the test set from Zhang *et al*. [58]. For pancreas sub-segmentation, we do not have a ground-truth or dataset for testing, because AbdomenAtlas 3.0 is the first public dataset to present pancreas sub-segments (head, body, and tail). However, our algorithm to sub-segment the pancreas closely follows radiologist-accepted standards (see Figure 5), and we asked radiologists to qualitatively evaluate our annotations.

4.3. RadGPT Enhances Human-made Reports

Human-made radiology reports often omit critical quantitative details—such as tumor volumes and attenuation (HU) values—compromising clinical decision-making. In our evaluation of 90 human reports from a UCSF, none reported organ or tumor volumes and only 63% measured all detected tumors. In contrast, our structured reports (RadGPT) consistently provide full, quantitative data. As shown in Table 4, while human-made reports measure volume and

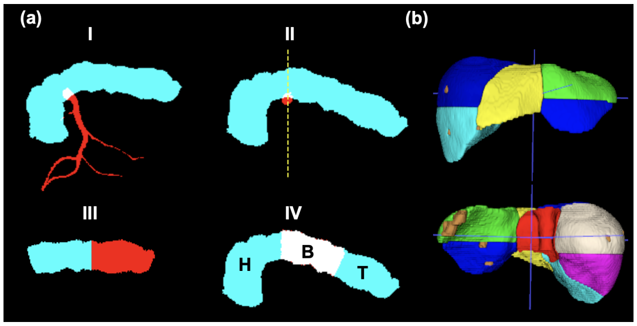

Figure 5. **Pancreas and liver sub-segments.** **(a)** RadGPT segments the pancreas based on radiology standards [48]. (I-II) the SMA separates the pancreas head (H) from the body (B), and (III) the remaining pancreas is divided at its midpoint into the body and tail (T). **(b)** Our liver sub-segmentation model achieved a DSC of 0.85 in segmenting eight liver sub-segments on a public test set [58]. Sub-segments are in different colors and tumors in brown. Sub-segments are essential for RadGPT to locate tumors.

HU in 0% of cases and capture all tumors in only 63% of cases, RadGPT achieves 100% for all metrics. This level of consistency streamlines clinical assessments and improves prognostic accuracy by ensuring every tumor is precisely measured [10, 39, 53].

In AbdomenAtlas 3.0, 240 CT scans include clinical notes, which lack quantitative tumor measurements; for instance, among 63 TotalSegmentator notes mentioning tumors, none provide such data—even though they report other findings like calcified arterial plaques. By merging these notes with our structured and narrative reports using an LLM, RadGPT generates 240 enhanced reports that integrate the notes' comprehensive clinical findings (covering 66 diagnoses) with precise tumor sizes (see Figure 12).

4.4. Discussion: End-to-End, Segmentation-based?

Table 2 shows RadGPT outperformed end-to-end VLMs. The unique design of RadGPT offers several **advantages** over end-to-end training. ***(i)*** Interpretable: RadGPT generates reports from tumor segmentation, allowing clinicians and developers to visualize and verify tumor locations and sizes in the CT. In contrast, errors in end-to-end methods are harder to diagnose and debug. ***(ii)*** Interactive: Our goal is creating a high-quality dataset of CTs and reports to drive innovation in report generation. As all algorithms make mistakes, a human-in-the-loop approach is key. Segmentation is a **safeguard**, allowing radiologists to ensure reports are correct, and easily transforming radiologist-revised masks into reports. ***(iii)*** Strong supervision: tumor segmentation AI has been a long-term focus of the research community, achieving high accuracy by leveraging precise per-voxel masks. Our benchmark shows that segmentation-assisted models can transfer this high accuracy to reports.

	Volume	HU	Diameters
Human-made	0%	0%	63%
RadGPT (ours)	100%	100%	100%

Table 4. **Comparison of human-made reports vs. RadGPT reports for 90 UCSF CTs.** Values indicate the percentage of reports containing tumor volume, HU and diameter measurements for all detected tumors. RadGPT reports provide more clinically relevant [10, 39, 53] quantitative details about tumors.

5. Conclusion & Future Work

Dataset curation and report-generation are inter-dependent. Developing image-report-mask methods requires image-report-mask datasets, but creating these datasets requires reliable methods and human-in-the-loop involvement. Our focus is creating high-quality image-report-mask datasets to support further methodological advancements.

AbdomenAtlas 3.0 is the first public dataset providing high-quality abdominal CT scans with reports and per-voxel tumor annotations, encompassing 9,262 CT scans from 138 institutions. It uniquely includes pancreas sub-segments, peripancreatic blood vessels, and pancreatic cancer stages—absent in existing public datasets. RadGPT transforms per-voxel annotations into structured reports using deterministic algorithms. These reports align with the accuracy of segmentations revised by radiologists in AbdomenAtlas 3.0. Additionally, RadGPT enables fully-automated report generation, surpassing existing approaches in detecting tumors. Together, AbdomenAtlas 3.0 and RadGPT bridge the gap between tumor segmentation and report generation, offering valuable resources and tools to advance AI in abdominal CT interpretation.

We are committed to expanding AbdomenAtlas 3.0 to include reports for more types of tumors. Additionally, we plan to host benchmarks using AbdomenAtlas 3.0 with two train/test splits. **IID Split:** Randomly sets aside 10% of the dataset for testing, where training and testing data come from the same institutions, following standard AI evaluation practices. Used in Table 2. **OOD Split:** Uses data from 23 unseen institutions (4,500 CT scans) for testing, providing a large test set to evaluate AI generalization to new environments. This benchmark will assess report generation models using standard text similarity metrics but will prioritize tumor detection sensitivity and specificity, enabled by our proposed LLM-based diagnostic evaluation.

Although 66 out of 240 fusion reports present diverse diagnoses, AbdomenAtlas 3.0 is cancer-centric. Cancer is a leading cause of death, and over 40% of medical imaging reports focus on cancer detection. Thus, AI-assisted report generation has the potential for significant impact. We hope our release of the cancer-centric AbdomenAtlas 3.0 will stimulate further AI advancements in the field.

Acknowledgments. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the National Institutes of Health (NIH) under Award Number R01EB037669. We would like to thank the Johns Hopkins Research IT team in IT@JH for their support and infrastructure resources where some of these analyses were conducted; especially DISCOVERY HPC. We thank the funding of Italian Institute of Technology and the HPC infrastructure at Italian Institute of Technology. P.R.A.S.B. thanks the funding from the Center for Biomolecular Nanotechnologies, Istituto Italiano di Tecnologia (73010, Arnesano, LE, Italy).

References

[1] Mahmoud M Al-Hawary, Isaac R Francis, Suresh T Chari, Elliot K Fishman, David M Hough, David S Lu, Michael Macari, Alec J Megibow, Frank H Miller, Koenraad J Mortele, et al. Pancreatic ductal adenocarcinoma radiology reporting template: consensus statement of the society of abdominal radiology and the american pancreatic association. *Radiology*, 270(1):248–260, 2014. 4, 5

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021. 3

[3] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):1–13, 2022. 3

[4] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024. 2, 3, 4, 6

[5] Pedro RAS Bassi, Sergio SJ Dertkigil, and Andrea Cavalli. Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization. *Nature Communications*, 15(1):291, 2024. 7

[6] Pedro RAS Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Yannick Kirchhoff, Maximilian Rokuss, Ziyan Huang, Jin Ye, Junjun He, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus H. Maier-Hein, Paul Jaeger, Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Yong Xia, Zhaohu Xing, Lei Zhu, Yousef Sadegheih, Afshin Bozorgpour, Pratibha Kumari, Reza Azad, Dorit Merhof, Pengcheng Shi, Ting Ma, Yuxin Du, Fan Bai, Tiejun Huang, Bo Zhao, Haonan Wang, Xiaomeng Li, Hanxue Gu, Haoyu Dong, Jichen Yang, Maciej A. Mazurowski, Saumya Gupta, Linshan Wu, Jiaxin Zhuang, Hao Chen, Holger Roth, Daguang Xu, Matthew B. Blaschko, Sergio Decherchi, Andrea Cavalli, Alan L. Yuille, and Zongwei Zhou. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? *Conference on Neural Information Processing Systems*, 2024. 7

[7] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019. 3, 5

[8] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *arXiv preprint arXiv:2406.06512*, 2024. 2, 4, 6, 12

[9] Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature medicine*, 29(12): 3033–3043, 2023. 4

[10] Z. Cao and et al. Volumetric versus linear measurements in lung cancer prognosis. *Annals of Surgical Oncology*, 26(12): 3757–3763, 2019. 8

[11] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 4

[12] Zeyuan Chen, Lingxi Xie, Xin Huang, Yu Rong, Shuguang Cui, and Bo Dai. Generating radiology reports via memory-driven transformer. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 231–240, 2020. 4

[13] Noel CF Codella, Ying Jin, Shrey Jain, Yu Gu, Ho Hin Lee, Asma Ben Abacha, Alberto Santamaria-Pang, Will Guyman, Naiteek Sangani, Sheng Zhang, et al. Medimageinsight: An open-source embedding model for general domain medical imaging. *arXiv preprint arXiv:2410.06542*, 2024. 2

[14] Errol Colak, Hui-Ming Lin, Robyn Ball, Melissa Davis, Adam Flanders, Sabeena Jalal, Kirti Magudia, Brett Marinelli, Savvas Nicolaou, Luciano Prevedello, Jeff Rudie, George Shih, Maryam Vazirabad, and John Mongan. Rsna 2023 abdominal trauma detection. `https://kaggle.com/competitions/rsna-2023-abdominal-trauma-detection`, 2023. Kaggle. 3

[15] Claude Couinaud. *Le foie: études anatomiques et chirurgicales*. Masson, 1957. 4

[16] Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12902–12915, 2024. 13

[17] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5

[18] Yasunari Fukuda, Daisaku Yamada, Hidetoshi Eguchi, Tomoki Hata, Yoshifumi Iwagami, Takehiro Noda, Tadafumi Asaoka, Koichi Kawamoto, Kunihito Gotoh, Shogo

Kobayashi, et al. Ct density in the pancreas is a promising imaging predictor for pancreatic ductal adenocarcinoma. *Annals of surgical oncology*, 24:2762–2769, 2017. 5

[19] F Gaillard et al. Radiopaedia: building an online radiology resource. European Congress of Radiology-RANZCR ASM 2011, 2011. 3

[20] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *CoRR*, 2024. 2, 4, 6, 12

[21] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–486. Springer, 2024. 2, 4, 6, 12

[22] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019. 3

[23] Nicholas Heller, Sean McSweeney, Matthew Thomas Peterson, Sarah Peterson, Jack Rickman, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Joel Rosenberg, et al. An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging., 2020. 3

[24] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 590–597, 2019. 7

[25] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 4

[26] Yuanfeng Ji, Hongliang Bai, Le Gu, Fucang Liu, Yingda Xia, Haofeng Lu, Zhennan Chen, Yunlong Gao, Hongyang Pan, Weilin Wang, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 111–120. Springer, 2022. 3

[27] Erin S. Kilpatrick and colleagues. Thresholds for low and high kidney volume in adult men and women. *National Center for Biotechnology Information*, 2018. 16

[28] Yoshihisa Kodama, Chaan S Ng, Tsung T Wu, Gregory D Ayers, Steven A Curley, Eddie K Abdalla, Jean Nicolas Vauthey, and Chusilp Charnsangavej. Comparison of ct methods for determining the fat content of the liver. *American Journal of Roentgenology*, 188(5):1307–1312, 2007. 5

[29] Shigeru Kondoh and colleagues. Ct imaging characteristics of the pancreas: Normal and abnormal findings. *PubMed*, 2018. 16

[30] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 3

[31] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, page 103285, 2024. 2, 3, 5, 17

[32] Wenxuan Li, Pedro RAS Bassi, Tianyu Lin, Yu-Cheng Chou, Xinze Zhou, Yucheng Tang, Fabian Isensee, Kang Wang, Qi Chen, Xiaowei Xu, et al. Scalemai: Accelerating the development of trusted datasets and ai models. *arXiv preprint arXiv:2501.03410*, 2025. 5

[33] Yuanhan Li, Jingyuan Chen, and Dinggang Shen. Knowledge-driven encode, knowledge-based decode: A novel pipeline for automatic chest x-ray report generation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 395–403, 2019. 4

[34] Y. Li, X. Wang, and Z. Zhang. Exploring cross-modal transformers for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages —, 2022. 4

[35] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, 2004. 4, 13

[36] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Tao Song, Xiaofan Zhang, Kang Li, Guotai Wang, and Shaoting Zhang. Word: Revisiting organs segmentation in the whole abdominal region. *arXiv preprint arXiv:2111.02403*, 2021. 3

[37] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[38] Jun Ma, Yuxin Zhang, Yuankai Gu, Haozhe Zhang, Zonghao Han, Jiancheng Dong, Yong Xie, Yefeng Zheng, and Yaowei Chen. Fast and low-gpu-memory abdominal organ segmentation from 3d ct images with nnformer and swinunetr. In *International Challenge on Abdominal Multi-Organ Segmentation from CT Images, FLARE 2022*, pages 83–94. Springer, 2022. 3

[39] A. McErlean and et al. Evaluation of tumor volume as a response metric in clinical trials and patient management. *Journal of Surgical Oncology*, 121(4):523–531, 2015. 8

[40] A.B. Miller, B. Hoogstraten, M. Staquet, and A. Winkler. Reporting results of cancer treatment. *Cancer*, 47(1):207–214, 1981. 5, 7

[41] Radiological Society of North America (RSNA). Radreport: Reporting templates. `https://www.radreport.org`, 2025. Accessed: 2025-01-06. 5

[42] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):1–9, 2020. 3

[43] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 556–564. Springer, 2015. 3

[44] Laura Schöckel, Gregor Jost, Peter Seidensticker, Philipp Lengsfeld, Petra Palkowitsch, and Hubertus Pietsch. Developments in x-ray contrast media and the potential impact on computed tomography. *Investigative radiology*, 55(9):592–597, 2020. 2

[45] J. Tang, P. Hu, and R. Qian. Multi-granular vision-language modeling for automated radiology reporting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages —, 2023. 4

[46] Andrew Taylor, William Dodds, Sandra Erickson, and Edward Stewart. Ct of acquired abnormalities of the spleen. *AJR American Journal of Roentgenology*, 157(6):1213–1219, 1991. 16

[47] Gianluca Tomasello, Michele Ghidini, Antonio Costanzo, Antonio Ghidini, Alessandro Russo, Sandro Barni, Rodolfo Passalacqua, and Fausto Petrelli. Outcome of head compared to body and tail pancreatic cancer: a systematic review and meta-analysis of 93 studies. *Journal of gastrointestinal oncology*, 10(2):259, 2019. 4

[48] Alexandre Triay Bagur, Paul Aljabar, Gerard R Ridgway, Michael Brady, and Daniel P Bulte. Pancreas mri segmentation into head, body, and tail enables regional quantitative analysis of heterogeneous disease. *Journal of Magnetic Resonance Imaging*, 56(4):997–1008, 2022. 5, 8

[49] Vanya V Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 547–556. IEEE, 2018. 3

[50] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 7

[51] Xiaosong Wang, Yifan Peng, Lu Li, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9049–9058, 2018. 4

[52] Mateusz Winder, Aleksander Jerzy Owczarek, Jerzy Chudek, Joanna Pilch-Kowalczyk, and Jan Baron. Are we overdoing it? changes in diagnostic imaging workload during the years 2010–2020 including the impact of the sars-cov-2 pandemic. In *Healthcare*, page 1557. MDPI, 2021. 2

[53] H.C. Woodruff and et al. The role of tumor volume in assessing treatment response in cancer imaging. *Cancer Imaging*, 15(1):5–12, 2015. 8

[54] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. 2, 4, 6

[55] Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, et al. The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, 2022. 4

[56] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*, 2022. 13

[57] A. Zhang, B. Chen, and C. Li. Hierarchical perception for chest x-ray image captioning and radiology report generation. In *European Conference on Computer Vision (ECCV)*, pages —, 2022. 4

[58] Xukun Zhang, Sharib Ali, Tao Liu, Xiao Zhao, Zhiming Cui, Minghao Han, Shuwei Ma, Jingyi Zhu, Yanlan Kang, Le Wang, et al. Robust and smooth couinaud segmentation via anatomical structure-guided point-voxel network. *Computers in Biology and Medicine*, 182:109202, 2024. 5, 7, 8

[59] Zongwei Zhou, Michael B Gotway, and Jianming Liang. Interpreting medical images. In *Intelligent Systems in Medicine and Health*, pages 343–371. Springer, 2022. 3