Appendix

Table of Contents

A. AbdomenAtlas 3.0 Dataset

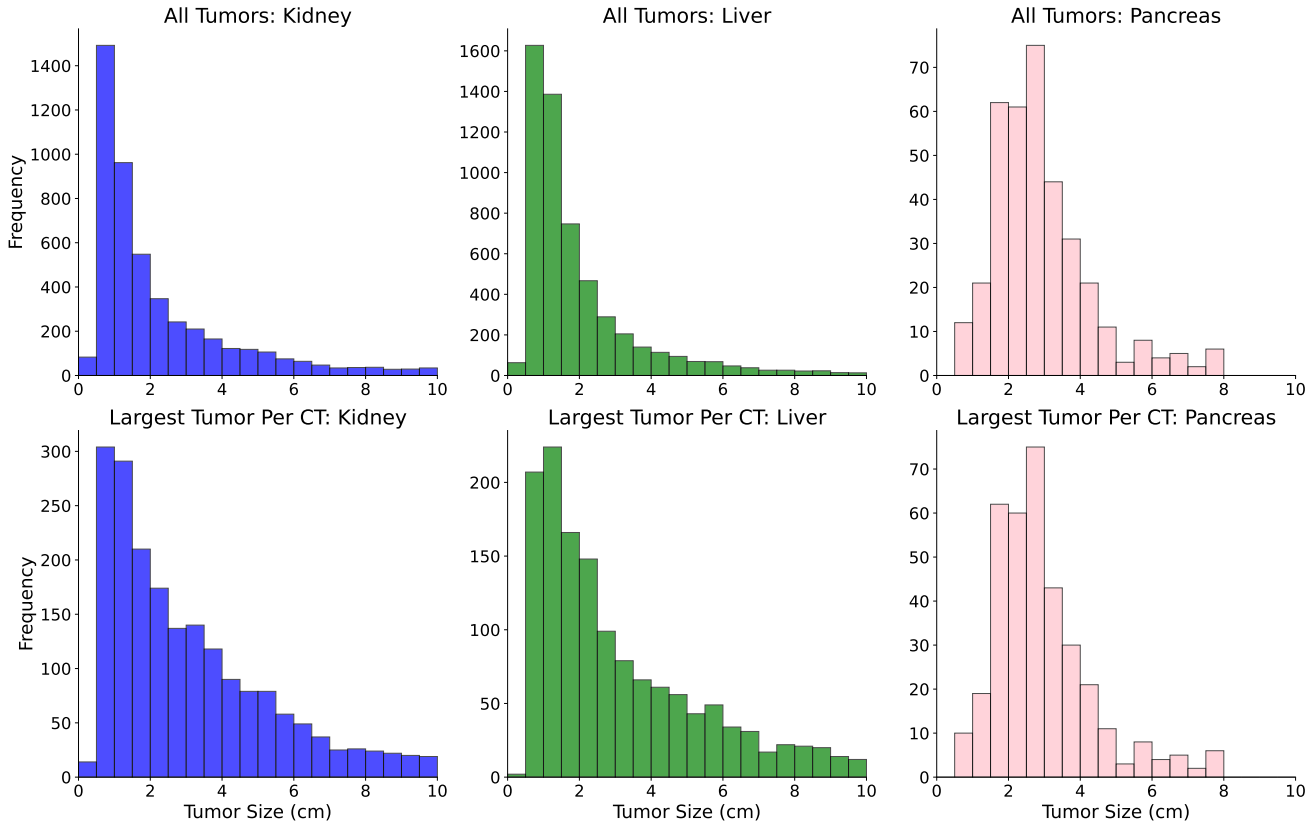

Figure 6. **Tumor size distribution in AbdomenAtlas 3.0. A large number of CT scans in AbdomenAtlas 3.0 present small tumors ($\leq$ 2 cm): 943.** The figure's top row shows histograms of all annotated tumors, while the bottom row focuses on the largest tumor in each organ. Notably, even considering only the largest tumor per organ, AbdomenAtlas 3.0 still includes a substantial number of small tumors ($\leq$ 2 cm): 504 for kidney, 358 for liver, and 81 for pancreas. These small tumor reports are crucial for training vision-language AI models to detect early-stage cancers, where identifying subtle abnormalities is critical for early detection and treatment.

A.1. Visualizations

A.1.1. Cancer Staging and Blood Vessels

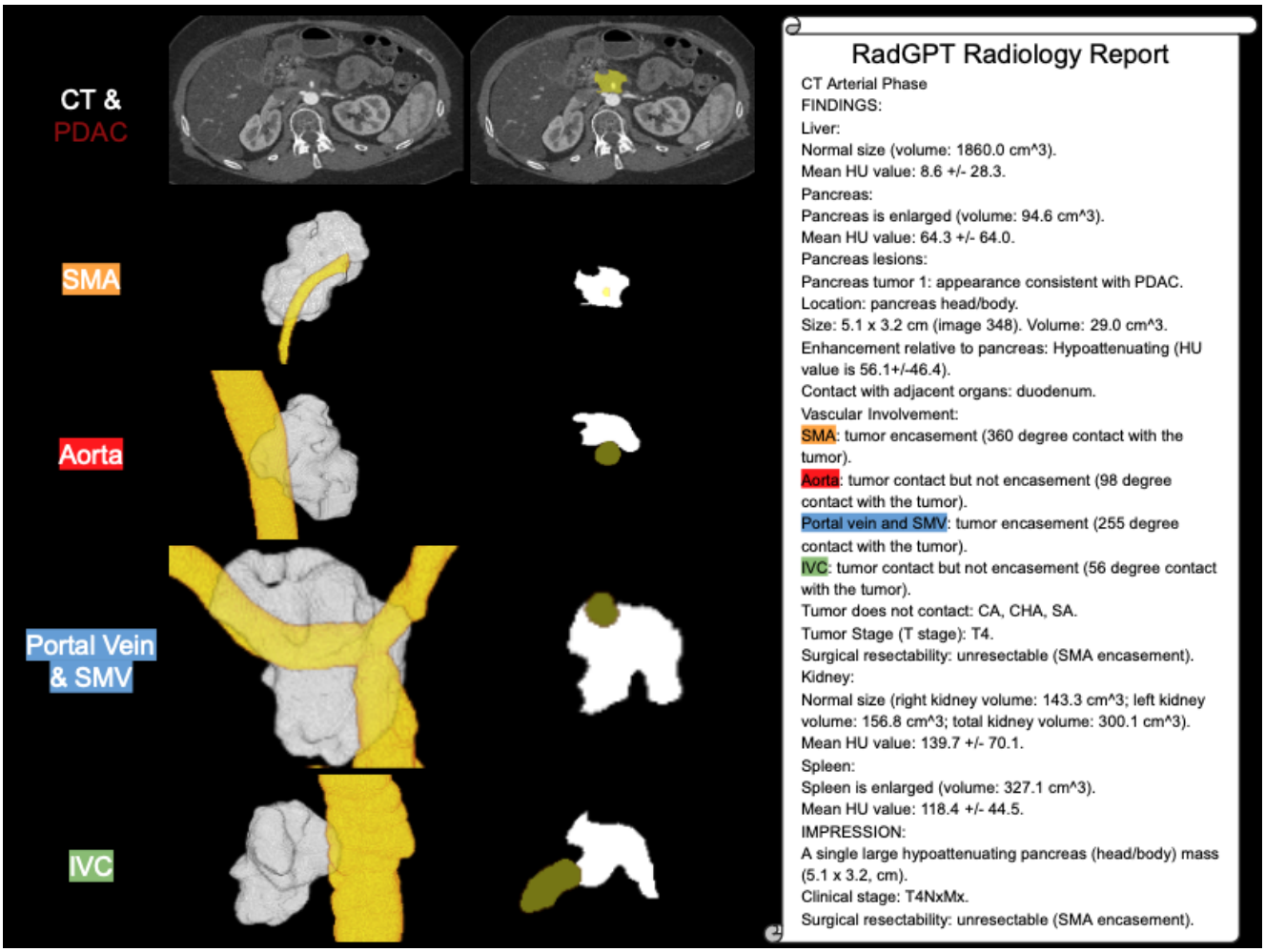

Figure 7. **Our pancreatic tumor (PDAC) staging report for a stage T4 tumor.** To determine the PDAC T stage, radiologists measure the tumor's size and evaluate its interactions with critical nearby blood vessels. RadGPT automatically replicates this process by utilizing per-voxel annotations of the PDAC and surrounding major blood vessels. The figure highlights these segmentations, and the report shows the angles of contact between the tumor and the blood vessels. In this case, the PDAC fully encases the superior mesenteric artery (SMA), which is a vital vessel supplying blood to the intestines. Surgical removal of a tumor encasing the SMA is not feasible because the artery cannot be reconstructed or bypassed without severe risk to the patient's survival. This involvement classifies the tumor as surgically unresectable and a stage T4 tumor.

A.1.2. Pancreas Sub-Segments

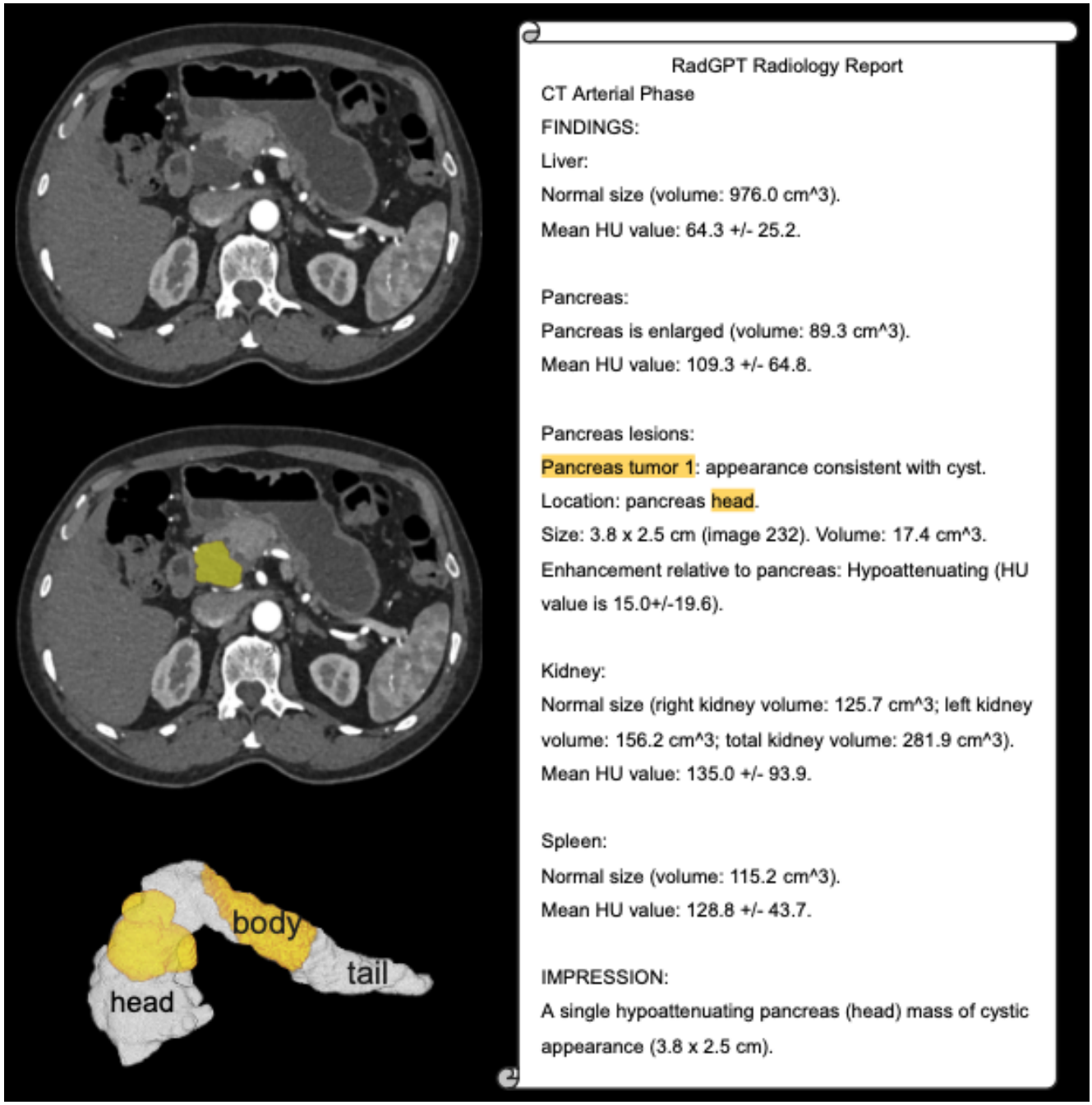

Figure 8. **CT scan with 2 pancreatic tumors (yellow), and illustration of pancreas sub-segmentation into head (white, left), body (yellow, middle) and tail (white, right).** RadGPT used the sub-segments to locate both PDAC tumors in the pancreas head. AbdomenAtlas 3.0 is the first to present pancreas sub-segments annotated per voxel. This information is crucial for writing radiology reports, as localizing pancreatic tumors in the pancreas head, body or tail is key for determining if the tumor can be surgically removed, and for tracking tumors in time.

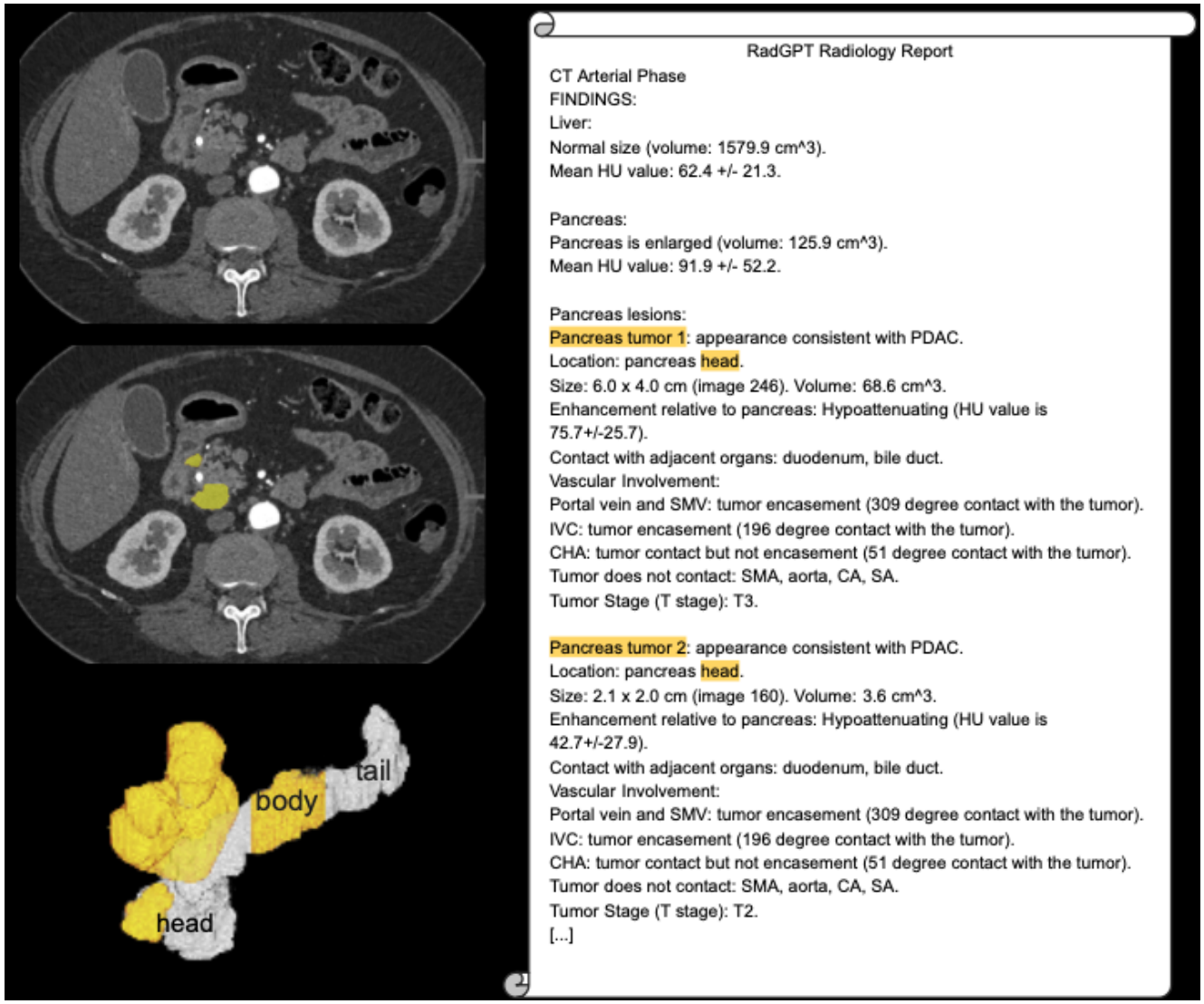

Figure 9. **CT scan with a pancreatic cyst (yellow), and illustration of pancreas sub-segmentation into head (white, left), body (yellow, middle) and tail (white, right).** RadGPT used the sub-segments to locate the cyst in the pancreas head. AbdomenAtlas 3.0 is the first to present pancreas sub-segments annotated per voxel. This information is crucial for writing radiology reports, as localizing pancreatic tumors in the pancreas head, body or tail is key for determining if the tumor can be surgically removed, and for tracking tumors in time.

A.1.3. Liver Sub-segments

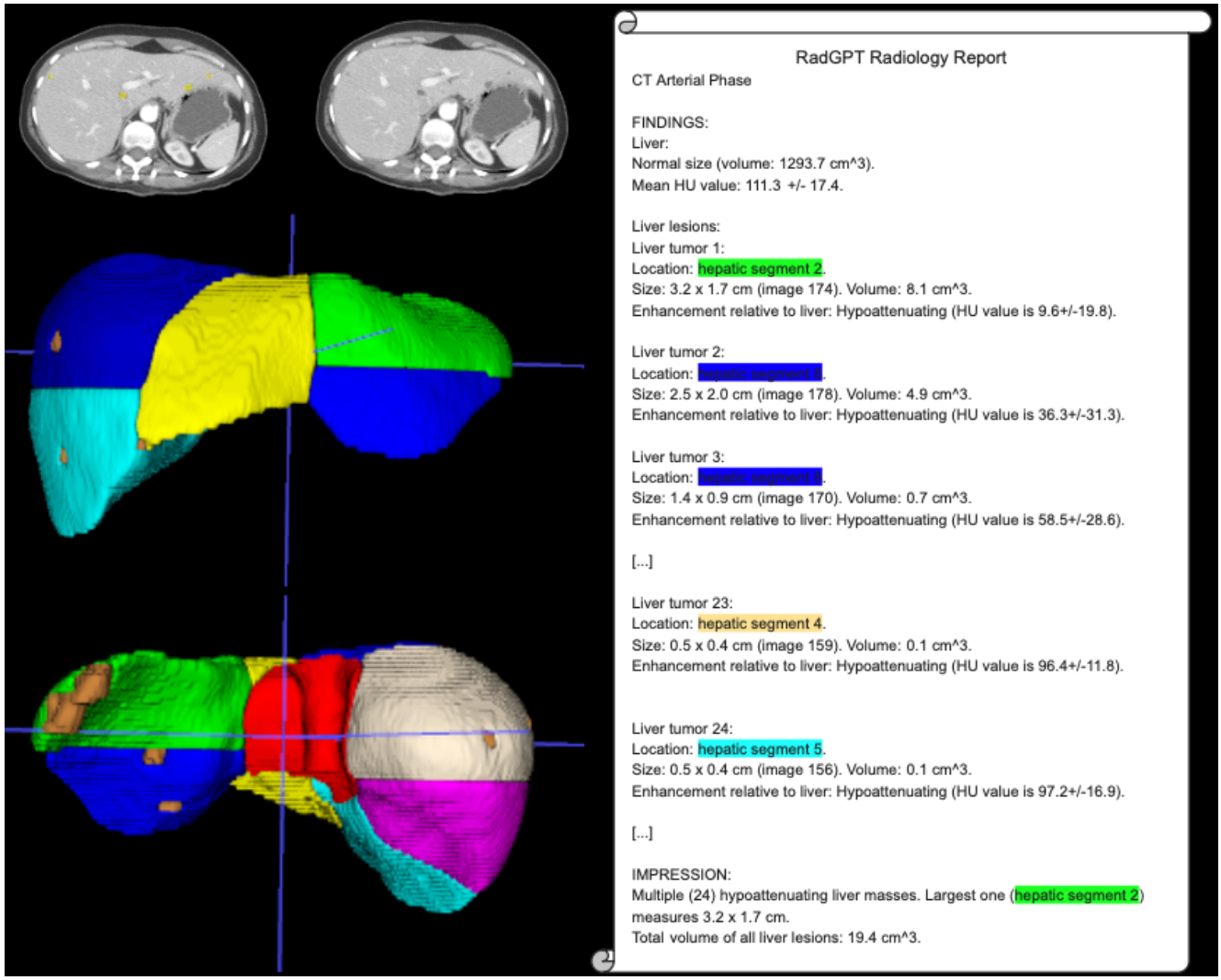

Figure 10. **CT scan with 24 liver tumors (brown), showing how we segment the liver into eight sub-segments for tumor localization.** Notably, unlike our report, most human-made reports would not describe 24 tumors in detail, due to the time required for this task. Liver sub-segments are functionally independent, and can be surgically removed without influencing nearby segments. Thus, localizing tumors into these segments is important for tracking tumors and for surgical planning.

A.1.4. Kidney Tumor Report

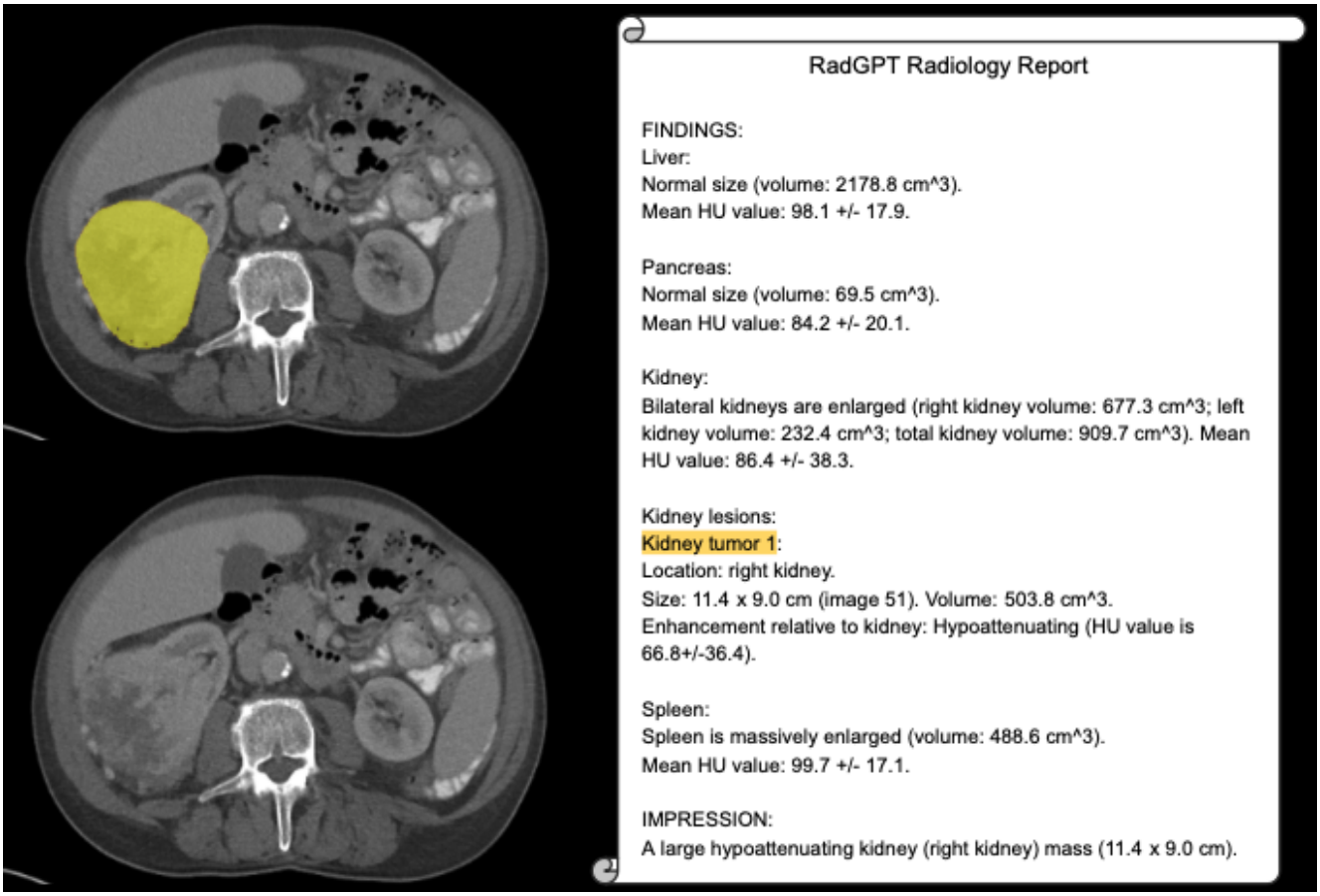

Figure 11. **CT scan showing a large kidney tumor (yellow) and our report.**

A.1.5. Enhanced Human Reports

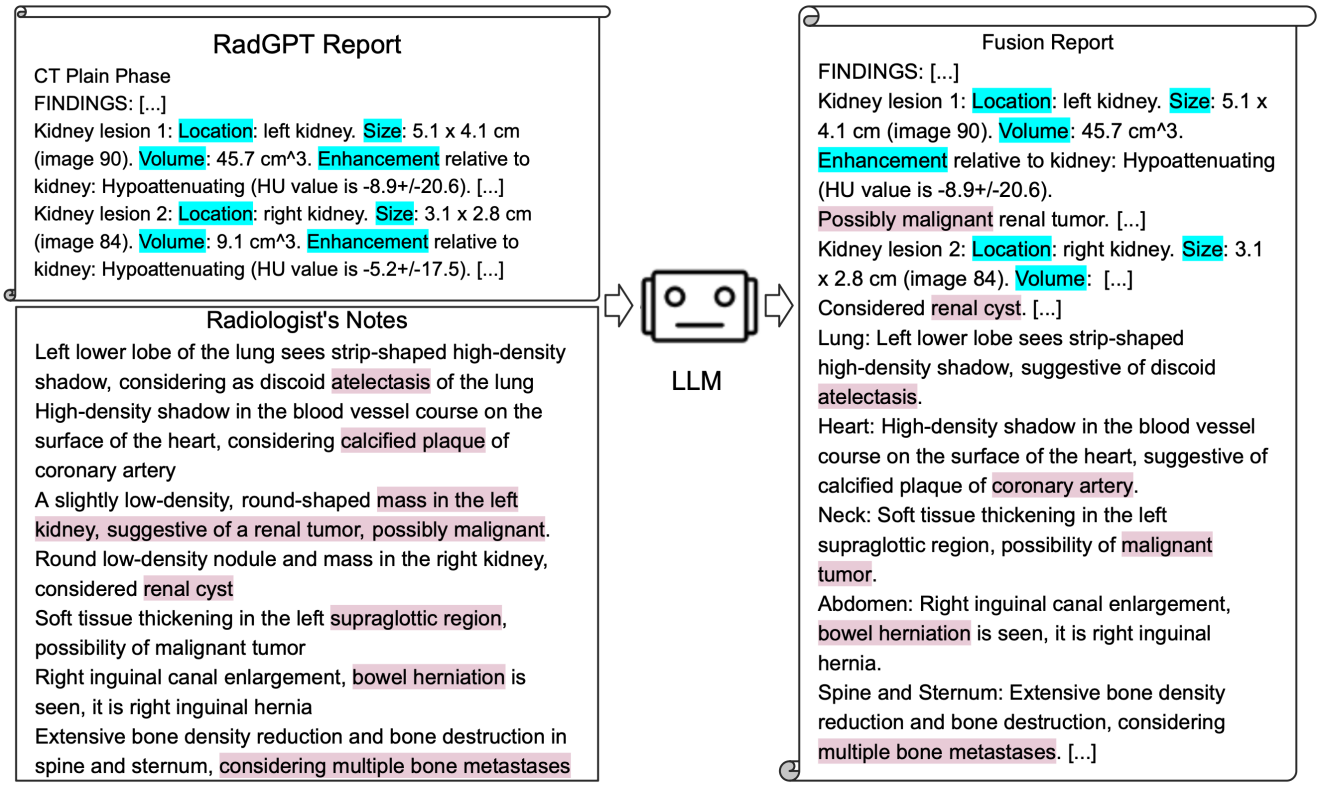

Figure 12. **In our enhanced human reports, LLMs combine detailed quantitative data from RadGPT's reports with the generality of human-made reports or clinical notes.** In the image, the LLM extracted kidney tumor sizes and volumes from the RadGPT report, while incorporating tumor type and non-cancer-related findings from the radiologist's notes. Reports created by RadGPT include detailed information often absent in human-made reports, such as tumor volumes and individual measurements for a large number of tumors when present. In contrast, human-made reports provide general findings that can expand RadGPT reports to many insights beyond cancer.

A.1.6. Narrative Reports

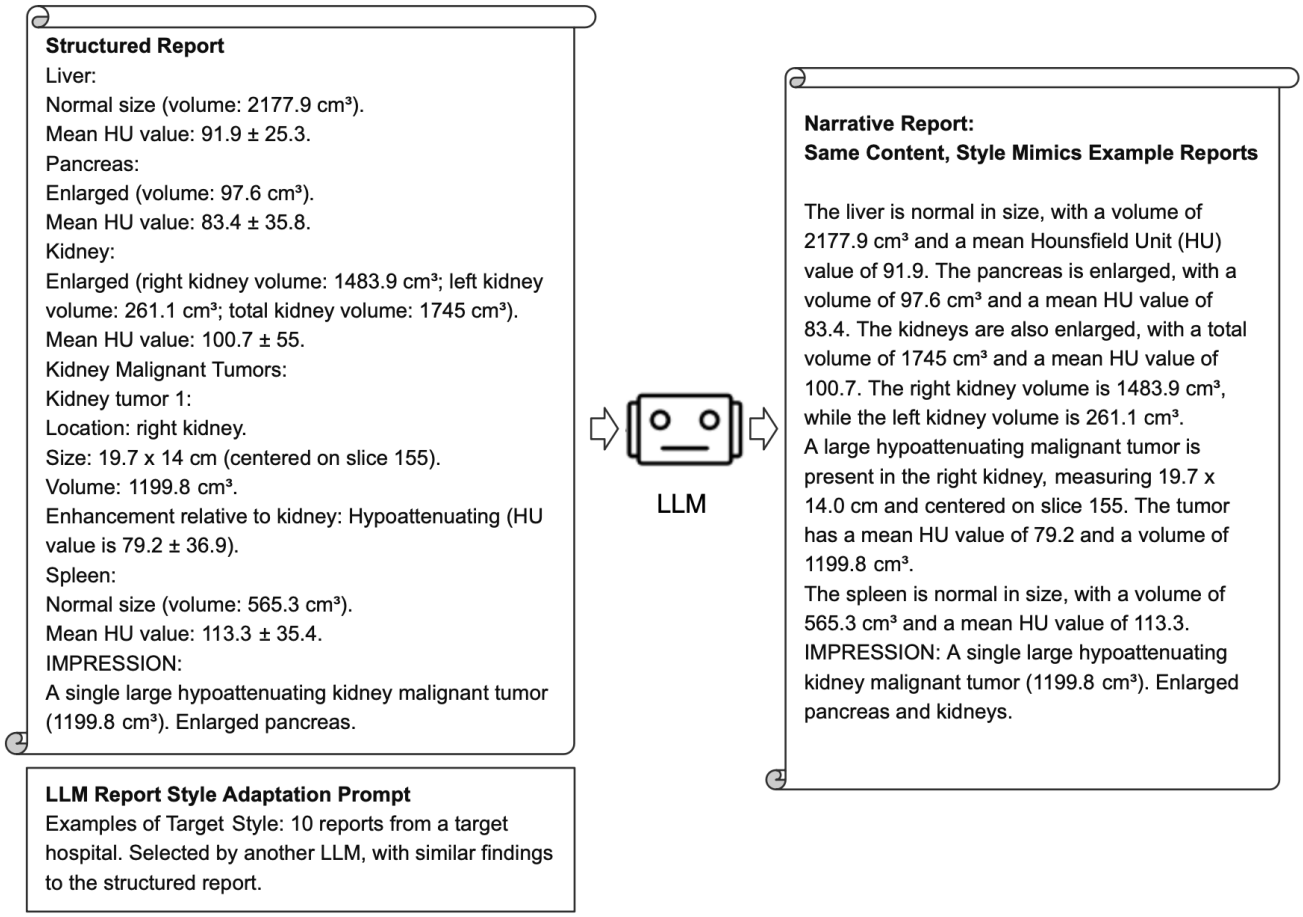

Figure 13. **Example of Narrative Report: we use LLM to convert structured reports into narrative reports that follow the writing style of a target hospital's.** The LLM receives 10 example reports from the hospitals as examples of style, and is instructed not to change the medical content of the structured report during style adaptation. Since reports targeting diverse abnormalities vary strongly in style, working and structure, we use another LLM to pre-classify the hospital's human-made reports into diagnostic categories (e.g., liver tumor). Thus, during style adaptation, we use as examples only reports that focus on the same diagnosis as the structured report. E.g., if the structured report mentions liver tumors, the examples also will concentrate on liver tumors.

A.2. Word Cloud

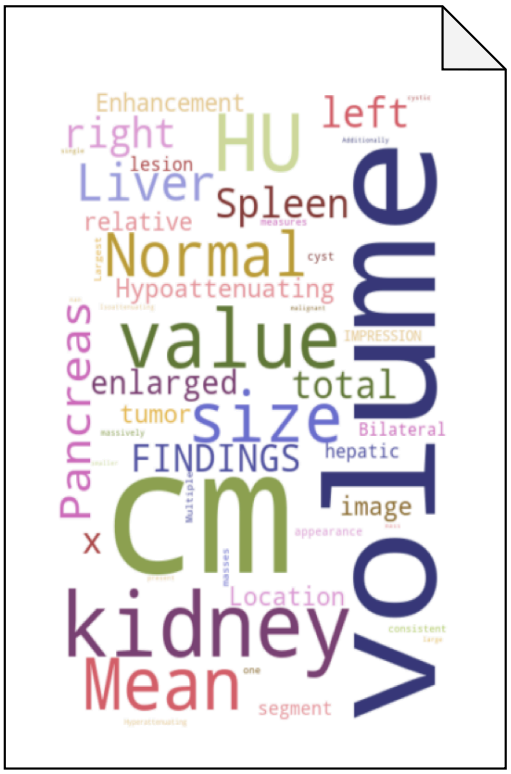

Figure 14. **Word cloud generated from the AbdomenAtlas 3.0 reports.** The size of each word reflects its frequency, highlighting the most common terms in the reports. The cloud provides insights about the reports, it clearly shows: their objective nature, focusing on measurements; the inclusion of volumes and HU values, which are usually absent from purely human-made reports; and the focus on cancer and tumor descriptions, with words like tumor, tumor, location, size, and enhancement.

B. Technical Details of RadGPT

Algorithm 1 Pancreas Sub-segmentation Using SMA

Erase parts of the SMA annotation below the pancreas annotation.
Perform PCA on a random subset of the pancreas voxels and rotate the pancreas around its center of mass, aligning its principal component with the x-axis. Rotate the SMA annotation together with the pancreas.
Project the SMA onto the x-axis; consider the x-plane at projection's midpoint as the boundary between pancreatic head and body.
For the remaining pancreas (excluding head), split body and tail at the x-axis midpoint, using another x-plane.
for each x-plane (slice) from tail to head **do**
    Identify connected components in the current slice.
    if first pancreas slice **then**
        Classify all components as body.
    else
        Classify components overlapping with the body in the previous slice as body; reclassify others as head. This is important for cases where the pancreatic head bottom crosses the SMA.
    end if
end for
Undo rotations and translations; save head, body, and tail segmentations.

Algorithm 2 WHO-based Tumors Size Measurement

Interpolate the tumor segmentation mask to a standard 1x1x1 mm spacing.
for each CT slice s containing tumor A **do**
    Extract the tumor borders by subtracting the tumor segmentation slice s by itself after binary erosion.
    Calculate the diameter D_s as the longest line between any two points in the tumor borders in s.
end for
Select the slice s_{max} with the largest diameter D_{max}.
In the selected slice s_{max}:
    Draw two lines L_1 and L_2 parallel to the diameter D_{max}.
    Set these two parallel lines L_1 and L_2 as far as possible from each other while touching the tumor borders.
    Calculate the distance d between lines L_1 and L_2.
Report the tumor size as $D_{max} \times d$, converting from mm to cm.

Algorithm 3 Automatic Tumor Staging

Make tumor borders overlap with vessels and organs
Apply binary dilation (3x3x3) on tumor mask.
for each vessel in {SMA, CHA, CA, SA} **do**
    if no overlap with tumor **then**
        Set `contact` = no and continue
    end if
    # Isolate main vessel branch
    for each slice along z-axis from top to bottom **do**
        Retain the largest connected component touching the previous slice's main component, or the largest if within the first 5% of slices.
    end for
    Apply binary erosion and dilation (5x5x5), overlap with original vessel segmentation, and retain the largest 3D component.
    # Check main branch contact with tumor
    if no overlap with tumor **then**
        Set `contact` = no and continue
    end if
    # Align vessel over x-axis and analyze contact with tumor
    Skeletonize main branch and align rotate volume, aligning principal component (PCA) with x-axis.
    for each x-coordinate along the x-axis **do**
        Check intersection with tumor; if none, continue
        Align 5mm vessel segment around x-axis using skeleton PCA and crop to 2.5mm
        # Calculate percentage of border contact with tumor to estimate contact angle (vessels are not perfectly round)
        Extract vessel borders for each slice and calculate border-tumor overlap percentage
        Compute contact angle as overlap percentage $\times$ 360; update `max_contact` for vessel if new maximum angle is found.
    end for
end for
# Define T stage based on vessel contact and tumor size thresholds
if `max_contact` for {SMA, CA, CHA} $\geq$ 180 **then**
    Stage = T4
else
    Determine stage by tumor size: T1a $\leq$ 0.5cm, T1b 0.5–1cm, T1c 1–2cm, T2 2–4cm, T3 $>$ 4cm
end if

B.1. Training CT2Rep & CT-CHAT & Merlin on AbdomenAtlas 3.0

We trained **CT2Rep** using only CT scans and structured reports, ignoring the per-voxel annotations in AbdomenAtlas 3.0. Our training strategy for CT2Rep closely followed the code and hyper-parameters published by the model authors [21]. Possibly, careful search of hyper-parameters and training algorithms for the abdominal region could improve the model's performance. We introduced minimal changes, needed to adapt the model to the abdominal region: we adopted sub-word tokenization to handle decimals frequently present in our reports; we standardized the CT spacing to 1.5 x 1.5 x 1.5 mm, a choice that reduces computational costs while facilitating tumor measurements by maintaining isotropy; to accommodate longer reports, we increased the model's maximum sequence length to 600; and, for hold-out validation (we used 30% of AbdomenAtlas 3.0 as the validation set), we used validation loss rather than sequential decoding and BLEU scoring, which significantly reduced validation time. These adjustments, while minimal, were designed to tailor the model for the unique challenges of abdominal CT report generation.

For **CT-CHAT**, we similarly trained the model using only CT scans and structured reports without using per-voxel annotations from AbdomenAtlas 3.0. While our general training approach again mirrored the original authors' published code and hyper-parameters [20], specific adaptations included standardizing the CT spacing to an isotropic 1.5 x 1.5 x 1.5 mm resolution, selecting a patch size of 20 in each dimension (x, y, and z), and ensuring consistency by center-cropping or padding scans to uniform dimensions of $300 \times 300 \times 600$ mm. Training was performed using four A100 GPUs for 20,000 iterations with a batch size of 16. Furthermore, we employed visual instruction fine-tuning identical to the CT-CHAT authors, using an attention pooling mechanism that reduced tokens generated by CT-CLIP to 256 via learned queries, which were then linearly transformed to match the hidden dimension of the Llama 3.1 8B model. Visual instruction fine-tuning proceeded for 100 epochs.

The training of **Merlin** also leveraged only CT scans and structured reports, again without incorporating per-voxel annotations from AbdomenAtlas 3.0. While closely aligning with the authors' original published code and hyper-parameters [8], Merlin required a distinct approach since only the pretrained volume encoder optimized for the abdominal region was available, without the report-generation weights. Consequently, we conducted visual instruction fine-tuning by applying a linear transformation to Merlin's encoded embeddings, mapping them to the hidden dimension of the Rad LLama2 7B model. This fine-tuning stage continued for 100 epochs, effectively adapting Merlin for abdominal CT report generation.

B.2. Segmentation Post-processing

Segmentation models can produce noise: voxels incorrectly labeled as tumors or organs. This may cause false positive cancer detections when RadGPT generates reports from nnU-Net or DiffTumor outputs. To address this, we propose a noise reduction algorithm (Alg. 4). Segmentation noise usually appears as small structures. Thus, we reduce it with binary erosion. Afterwards, to restore the original shape of true tumors and organs, we applied binary dilation followed by a voxel-wise AND with the original tumor segmentation. To further avoid false positives, we perform organ-wise thresholding: we only consider an organ has tumors if the total volume of its tumor voxels is above a small threshold, defined to maximize per-class F1-Score on a validation dataset. For our results section, RadGPT thresholds are: 1 mm^3 in the pancreas, 150 mm^3 in the kidneys, 100 mm^3 in the liver, and 50 mm^3 for metastases. Figure 15 shows specificity and sensitivity for multiple thresholds. Algorithm 4 and thresholding are not necessary when we generate AbdomenAtlas 3.0 reports from radiologist revised segmentations or ground-truth segmentation masks. However, it is recommended when using RadGPT without human revision (Figure 2). Figure 15 displays performance variation for diverse thresholds.

Algorithm 4 Segmentation Noise Reduction

1: Copy the segmentation output.
2: Apply binary erosion to the segmentation to erase small structures, considered noise. We use a 3x3x3 structuring element, erasing any structure smaller than a 3x3x3 cube.
3: Perform binary dilation on the eroded segmentation. We use a 4x4x4 structuring element.
4: Apply a voxel-wise AND operation between the original mask (before erosion) and the dilated mask, recovering the shape of structures not removed by the binary erosion.

B.3. RadGPT Enables Diagnostic Evaluation

In Table 5, we use standard text similarity metrics (common in LLM evaluation) and RadGraph-F1 to evaluate the reports generated by the AI. RadGPT achieves the highest scores in BLEU, METEOR, and ROUGE. These results align with the superiority of RadGPT in our diagnostic evaluation (Table 2). Thus, diagnostic accuracy may improve LLM metrics. However, BERT and RadGraph-F1 are not aligned with the diagnostic accuracy in Table 2—RadGPT has considerably superior diagnostic accuracy for cancer (Table 2), but it does not have the highest BERT and RadGraph-F1 scores. Moreover,

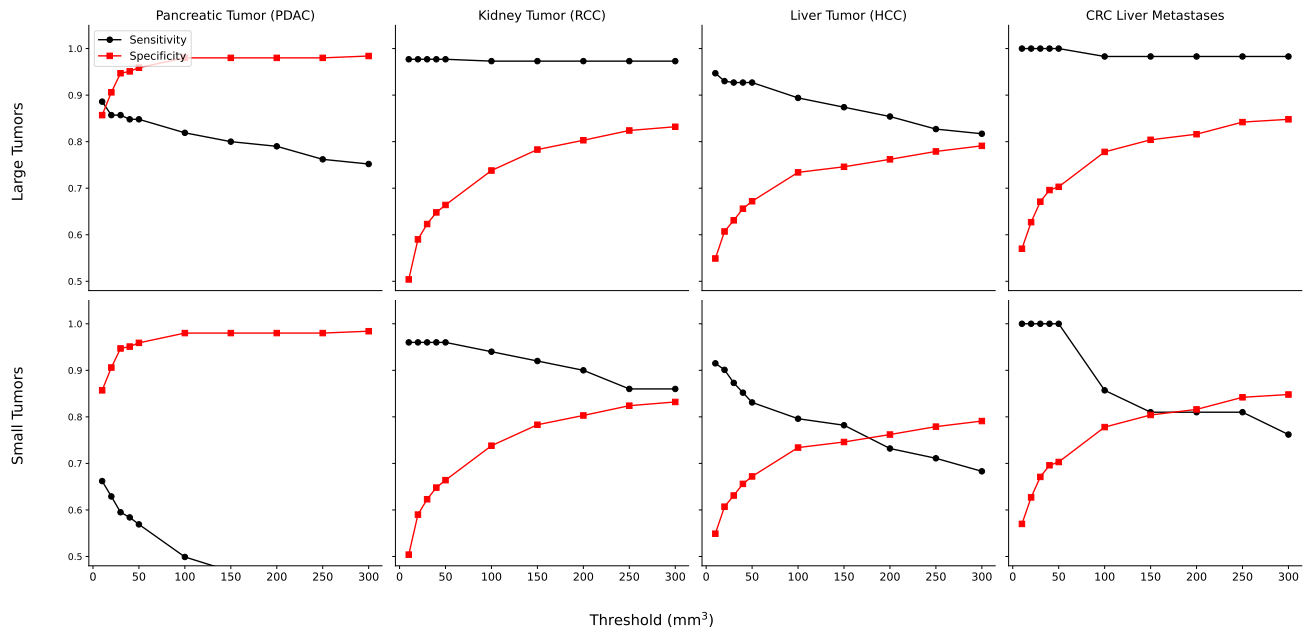

Figure 15. **Tumor detection sensitivity and specificity for RadGPT with diverse thresholds.** Evaluation performed on a private dataset from a hospital never observed during training (UCSF), detailed in Table 2.

model	BLEU	METEOR	ROUGE-1	R.-2	R.-L	BERT	RadGraph-F1
CT2Rep	0.26	8.27	12.73	0.95	7.46	45.19	3.01
CT-CHAT	0.15	7.19	10.90	0.54	6.00	45.63	5.69
M3D	0.02	5.05	12.51	1.80	7.32	44.05	1.69
Merlin	0.17	7.30	11.63	0.72	6.38	45.44	**6.20**
RadFM	0.10	6.90	16.14	2.41	9.26	**46.40**	2.36
RadGPT-S	0.72	8.82	13.56	1.22	7.64	45.62	4.15
RadGPT-N	**0.74**	**10.47**	**22.77**	**2.64**	**11.89**	45.41	2.92

Table 5. **Report style impacts standard evaluation metrics.** Testing on the same unseen hospital as in Table 2 (UCSF), we evaluate the fully-automated reports from RadGPT with text similarity metrics and RadGraph-F1[56] (using RadGraph-XL [16], which covers the abdomen region). RadGPT narrative (N) and structured (S) reports differ in style only, having the same sensitivity and specificity for tumor detection. However, by mimicking just the style of the test hospital (UCSF, §3.2), our narrative reports (RadGPT-N) achieved considerably higher METEOR and ROUGE scores. In contrast, our proposed evaluation metric (Table 2) only evaluates the diagnosis in the reports and it is not sensible to style variations.

NLP metrics are affected by the style of the report. Table 5 includes structured and narrative reports by RadGPT. They contain the same diagnoses and details, but different style (§3.2), with the narrative reports mimicking the style of the test hospital. In Table 5, ROUGE [35] was the metric most affected by the style variation. The results show that standard LLM metrics are influenced by style, but the extent of this influence varies between metrics. Conversely, our proposed LLM-based sensitivity and specificity metrics are only influenced by the diagnostic accuracy of the reports. In addition, sensitivity and specificity provide clinicians with an objective and easily interpretable evaluation of AI-made reports, objectively measuring the clinical usefulness of a vision-language model.

B.4. LLM Prompts

B.4.1. Style Adaptation

Or prompt for style adaptation is the following:

”””

You are provided with a **structured radiology report** and n other radiology reports that have different writing styles compared to the structured report.

Task:

Please **paraphrase** the structured report to match the writing style of the other reports.

Important Guidelines:

1. **Do Not Alter Medical Information:** Do not change, add, or remove any medical details such as tumor measurements, types, or locations. You may remove HU values.
2. **Maintain Original Meaning:** Ensure that the rephrased report conveys the same information as the original structured report.
3. **Match Writing Style:** Adapt the language, tone, and structure to align with the provided example reports.
4. **Begin your report text with #start and finish it with #end.**
5. **Provide justification:** Go through all medical findings in your rephrased report (e.g., tumor size, no evidence of metastasis) and show where the information comes from in the structured report. Justification should come after #end.
6. **Pay attention to the Example Reports:** Your writing style must be consistent with the examples.
7. **Organization must match:** If the examples have an *Impressions* and *Results* section, you must add them. If the example reports talk about all abdominal organs in a single paragraph, you must do so too. You may skip sections you cannot fill due to lack of information, like patient history.
8. **Do not add new findings:** If the structured report does not mention the presence or absence of a medical condition (e.g., metastases), you must NOT include it in your rephrased report.
9. **Keep coherence:** Avoid going back and forth between medical findings or organs. For example, do not talk about the size of a pancreatic tumor, then mention the liver, and then go back to pancreatic findings. Keep the information about each organ together.
10. **Always include an impressions section with the most important findings.**

Example of Rephrasing:

Structured Report:

PDAC 1: Pancreatic body/tail. Hypoattenuating pancreas PDAC measuring 6.0 x 3.4 cm (centered on slice 356). Its mean HU value is 39.17 +/- 29.65, and its volume is 27.519 cm^3.

Paraphrased Report:

#start

The patient has a pancreatic adenocarcinoma located in the body and tail of the pancreas, measuring 6.0 x 3.4 centimeters (image slice 356). The tumor is hypoattenuating and has a volume of 27.519 cm^3.

#end

Justification:

a. **Tumor Type:** Maintained as ”pancreatic adenocarcinoma”, originally ”PDAC”.
b. **Location:** Preserved as ”body and tail of the pancreas”, originally ”Pancreatic body/tail”.
c. **Measurements:** Kept as ”6.0 x 3.4 centimeters”, originally ”measuring 6.0 x 3.4 cm”.
d. **Imaging Slice:** Retained as ”image slice 356”, originally ”centered on slice 356”.
e. **Attenuation:** Maintained as ”hypoattenuating”, originally ”Hypoattenuating pancreas PDAC”.
f. **Volume:** Kept as ”27.519 cm^3”, originally ”volume is 27.519 cm^3”.

Note: Removed mean HU value as per guidelines.

”””

Example Reports (Target Style): {examples}
Structured Report to Paraphrase: {structured_report}

B.4.2. Enhancing Human Reports

Our Report Fusion prompt is:

”””

You are provided with a CT scan **structured radiology report** and notes written by a radiologist, about the same CT scan.

Your task is to identify any information in the notes that is not already included in the structured report and add it to the appropriate sections of the report. Please follow these guidelines:

1. **Do not remove** any existing information from the structured report. However, you may improve the report's details using **only** relevant information from the notes.
2. **Avoid adding any new findings** not already mentioned in either the notes or the structured report.
3. **Maintain the report's structure.** Carefully place new information in the correct sections inside ”FINDINGS”, considering which organ the information mentions. For instance, if the notes mention ”cirrhosis,” add it to the **”Liver”** section under **”FINDINGS”**.
4. **Add new sections if necessary.** If the notes refer to an organ not covered in the structured report, create a new section for it. If the notes mention patient metadata (e.g., sex and age), you may add it to the beginning of the report.
5. **Update the IMPRESSION section if needed.** Besides the FINDINGS, include any critical information from the notes in the report's **IMPRESSION** section, summarizing or rephrasing it. Do not add new sections if the notes do not provide concrete information for them.
6. **Use consistent terminology.** If possible, make the terminology in the sentences you add to the report match the terminology in the original structured report.
7. **Begin your report text with #start and finish it with #end.**
8. **Provide justification.** Explain where in the report you added each piece of information from the notes. Also, explain why other information in the report was not removed or altered.
9. **Do not** write non-informative sentences such as ”Patient metadata: Not available in the provided notes” or ”Sex: Not specified.”

The notes are as follows:

{clinical_info}

The current structured report is:

{structured_report}

”””

B.4.3. Labeling/Report Evaluation

Our prompt is:

Instructions: Discover if the CT scan radiology report below indicates the presence of liver tumors, pancreas tumors, or kidney tumors. Output labels for each of these categories: **yes** to indicate tumor presence, **no** for tumor absence, and **U** for uncertain tumor presence.

Example: liver tumor presence=yes; kidney tumor presence=U; pancreas tumor presence=no.

Answer with only the labels, do not repeat this prompt.

Follow these rules for interpreting radiology reports:

1. **'Unremarkable'** means that an organ has no tumor.
2. Multiple words can describe tumors. Check both the **findings** and **impressions** sections of the report (if present) to understand if an organ has tumors. Some words include: metastasis, tumor, tumor, mass, cyst, neoplasm, growth, cancer, index tumor in cancer patients, and tumors listed as oncologic findings.
3. Consider any tumor, hyperdensity, or hypodensity a tumor, unless the report explicitly states otherwise. Many conditions are not tumors and should not be interpreted as such unless a tumor is also reported along with the disease. Examples include:
   - **Liver conditions:** Hepatitis, Cirrhosis, Fatty Liver Disease (FLD), Liver Fibrosis, Hemochromatosis, Primary Biliary Cholangitis (PBC), Primary Sclerosing Cholangitis (PSC), Wilson's Disease, Liver Abscess, Alpha-1 Antitrypsin Deficiency (A1ATD), Steatosis, Granulomas, Cholestasis, Budd-Chiari Syndrome (BCS), Transplant, Gilbert's Syndrome, ulcers, wounds, infections, inflammations, and scars.
   - **Kidney conditions:** Stents, inflammation, postinflammatory calcification, transplant, Chronic Kidney Disease (CKD), Acute Kidney Injury (AKI), Glomerulonephritis, Nephrotic Syndrome, Polycystic Kidney Disease (PKD), Pyelonephritis, Hydronephrosis, Renal Artery Stenosis (RAS), Diabetic Nephropathy, Hypertensive Nephrosclerosis, Interstitial Nephritis, Renal Tubular Acidosis (RTA), Goodpasture Syndrome, and Alport Syndrome.
   - **Pancreas conditions:** Pancreatitis, Pancreatic Insufficiency, Cystic Fibrosis (CF), Diabetes Mellitus (DM), Exocrine Pancreatic Insufficiency (EPI), Pancreatectomy, and Pancreatic Pseudocyst.
4. Examples of specific tumor names include:
   - **Liver:** Hepatic Hemangioma (HH), Focal Nodular Hyperplasia (FNH), Bile Duct Adenoma, Simple Liver Cyst (SLC), Hepatocellular Carcinoma (HCC), Cholangiocarcinoma (CCA), Hepatic Adenoma (HA), Mucinous Cystic Neoplasm (MCN).
   - **Pancreas:** Serous Cystadenoma (SCA), Pancreatic Ductal Adenocarcinoma (PDAC), Mucinous Cystadenocarcinoma (MCC), Mucinous Cystadenoma (MCA), Intraductal Papillary Mucinous Neoplasm (IPMN), Solid Pseudopapillary Neoplasm (SPN), Pancreatic Neuroendocrine Tumor (PNET).
   - **Kidney:** Renal Oncocytoma (RO), Angiomyolipoma (AML), Simple Renal Cyst, Bosniak IIF Cystic Tumor, Renal Cell Carcinoma (RCC), Transitional Cell Carcinoma (TCC), Wilms Tumor, Cystic Nephroma (CN), Multilocular Cystic Renal Neoplasm of Low Malignant Potential (MCRNLMP), Hydronephrosis, Allograft.
5. Consider any benign (e.g., cyst) or malignant tumor as a tumor. Thus, any type of cyst is a tumor.
6. Organs never mentioned in the report have no tumors.
7. Do not assume a tumor is uncertain unless it is explicitly reported as uncertain. Many words can describe uncertainty, such as: ill-defined, too small to characterize, nonspecific, and uncertain. Reports may express uncertainty about tumor type (e.g., cyst or hemangioma) but still confirm it is a tumor—in this case, consider the tumor a tumor.
8. Organs with no tumor but other pathologies should be reported as **no**.

B.5. Organ size standards

Our standards for considering organs as large are based on widely accepted thresholds in radiological and anatomical studies. For the spleen, we consider volumes greater than 314.5 cm^3 as large and over 430.8 cm^3 as massive, based on thresholds provided by Taylor et al. [46]. For the kidneys, a volume exceeding 415.2 cm^3 for men is considered large, with the threshold adjusted for individual kidneys (half of the total volume) [27]. Similarly, a liver volume exceeding 3000 cm^3 is deemed large, which represents an upper limit for larger individuals, such as a 150 kg man, and highly depends on factors like weight and sex. For the pancreas, volumes above 83 cm^3 are classified as large, as per imaging standards discussed by Kondoh et al. [29].

When size standards depend on variables like weight or sex, we apply thresholds suitable for larger individuals to ensure comprehensive assessments. This approach minimizes the risk of underestimating organ size variations in diverse populations.

C. Revisions by Radiologists

In AbdomenAtlas 3.0, organ segmentation masks were created and verified by radiologists through an efficient human-in-the-loop approach [31]. Conversely, the tumor segmentation masks were suggested by AI and radiologists individually verified and corrected them. To verify reports efficiently, radiologists first ensured that structured reports correctly described the already revised per-voxel tumor annotations. This confirmation was key to ensure our deterministic algorithms worked correctly. Then, to verify our narrative and enhanced human reports, we first used our double-check procedure: an LLM (Llama 3.1 70B) extracted tumor information from the narrative / human enhanced reports and checked if it matched the information in the corresponding structured reports. For cases of mismatch, we prompted the LLM to correct the narrative / human enhanced reports and repeated the double-check. Any remaining mismatch was sent to radiologists. Mismatches were also analyzed to improve our prompts. E.g., radiologists identified that a few narrative or enhanced reports introduced findings absent from the source structured report or human report. To correct this, we started prompting the LLM to justify each finding by quoting sentences in the source structured report or human report. After the double check, we organized structured, narrative and enhanced human reports in a table and radiologists could quickly compare them, confirming they had consistent medical findings.

D. Detailed Tumor Statistics

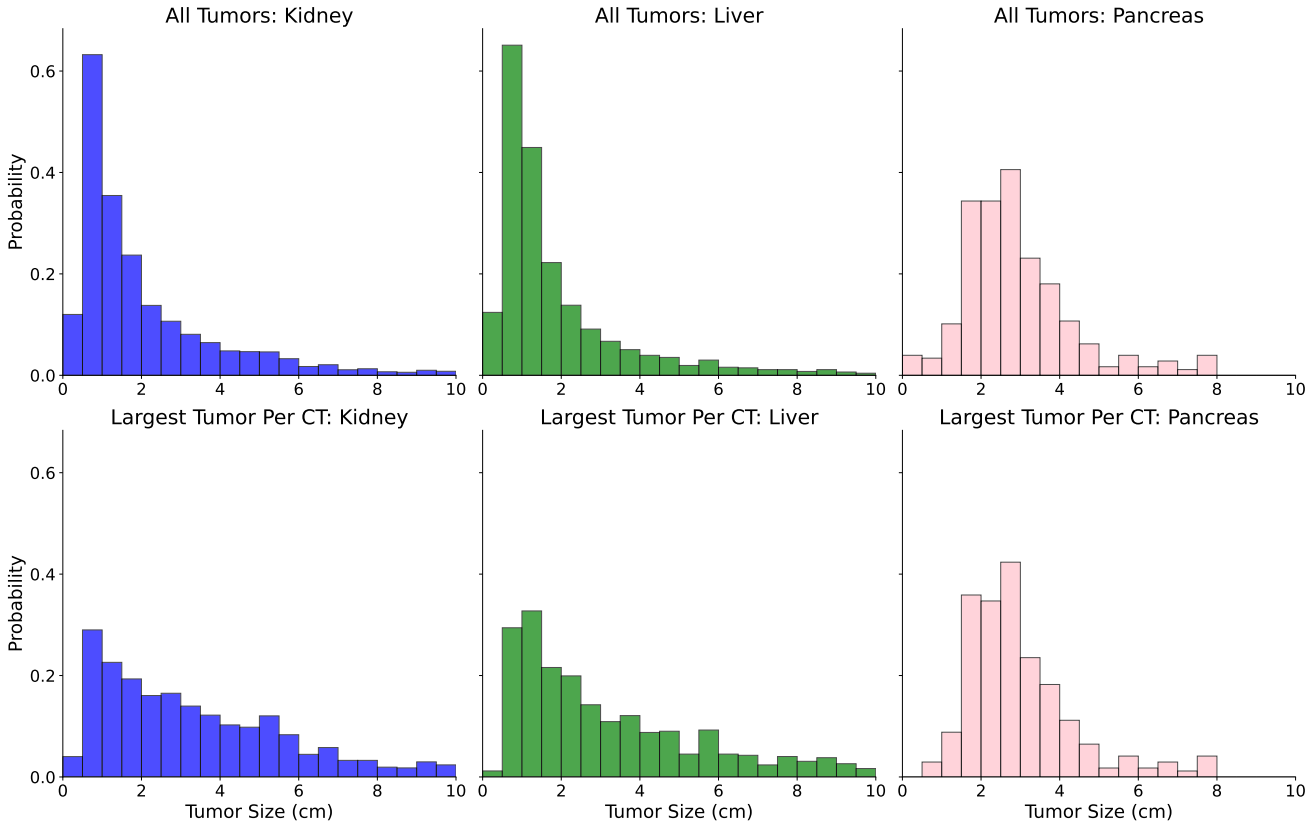

Figure 16. **Tumor size distribution in AbdomenAtlas 3.0. A large proportion of the CT scans, 35%, presents small tumors only ($\leq$ 2 cm).** The figure's top row shows histograms of all annotated tumors, while the bottom row focuses on the largest tumor in each organ. Notably, even considering only the largest tumor per organ, the proportion of small tumors ($\leq$ 2 cm) is large in AbdomenAtlas 3.0: 35.59% for kidney, 38.25% for liver, and 23.68% for pancreas. These small tumor reports are vital for training vision-language AI models to detect early-stage cancers, where identifying subtle abnormalities is critical for early cancer detection and treatment.

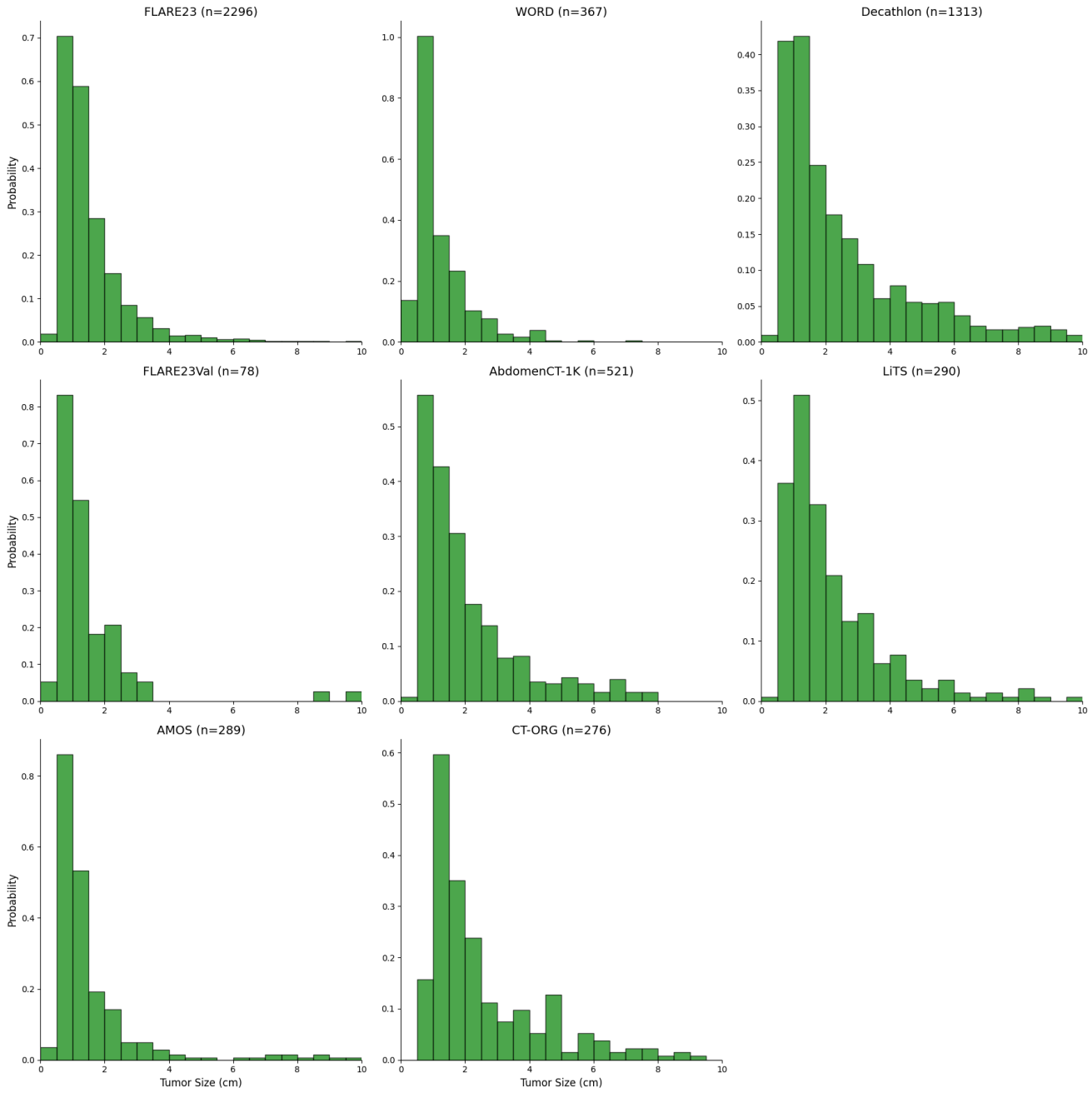

Figure 17. **Tumor size probability distribution for liver tumors across all datasets in AbdomenAtlas 3.0.** Each subplot represents a dataset with at least three tumor occurrences. The x-axis shows tumor size (cm), and the y-axis represents the probability of tumors within each size range. The figure highlights the variability in tumor sizes annotated across datasets, and the significant presence of small tumors.

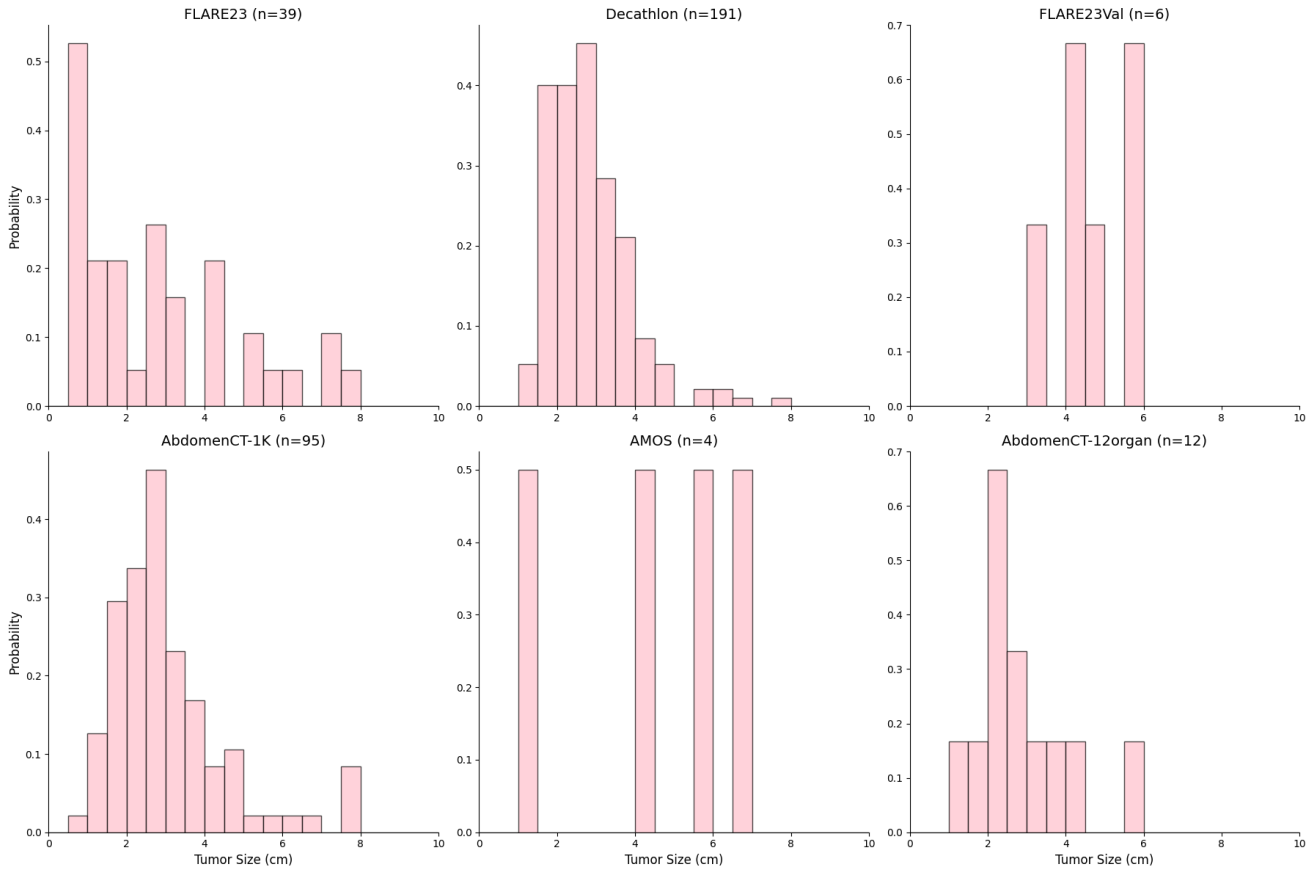

Figure 18. **Tumor size probability distribution for pancreas tumors across all datasets in AbdomenAtlas 3.0.** Each subplot represents a dataset with at least three tumor occurrences. The x-axis shows tumor size (cm), and the y-axis represents the probability of tumors within each size range. The figure highlights the variability in tumor sizes annotated across datasets, and the significant presence of small tumors.

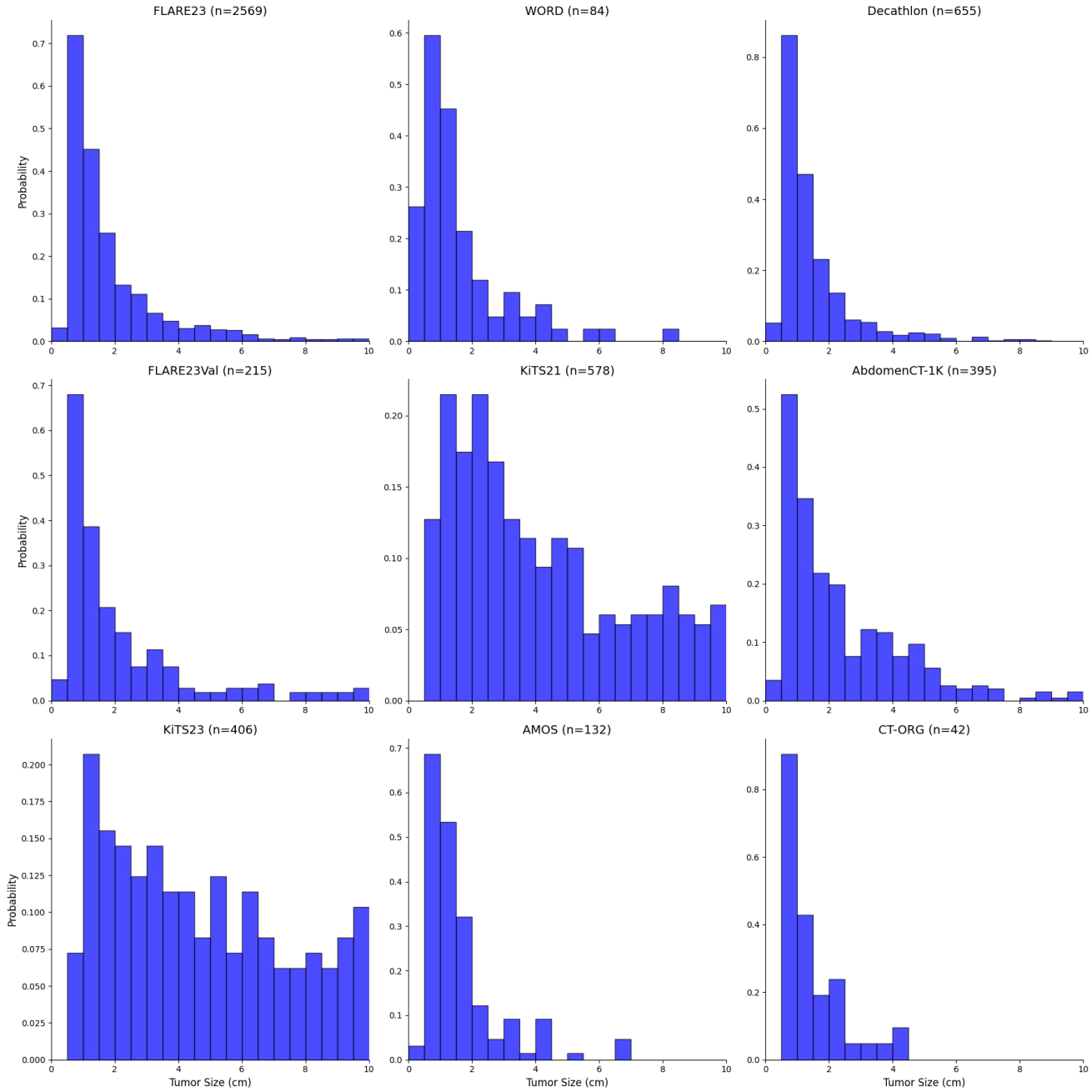

Figure 19. **Tumor size probability distribution for kidney tumors across all datasets in AbdomenAtlas 3.0.** Each subplot represents a dataset with at least three tumor occurrences. The x-axis shows tumor size (cm), and the y-axis represents the probability of tumors within each size range. The figure highlights the variability in tumor sizes annotated across datasets, and the significant presence of small tumors.

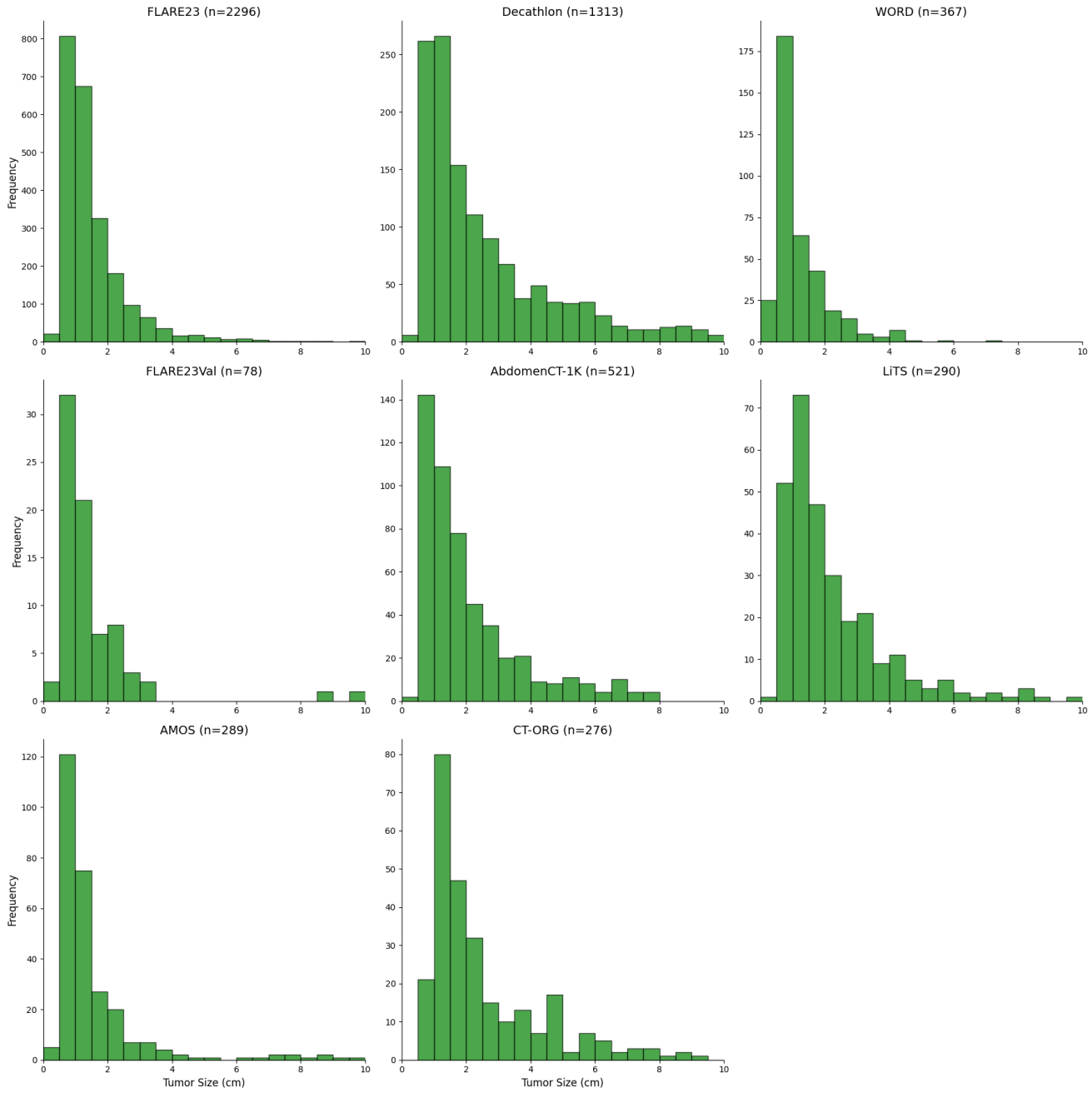

Figure 20. **Tumor size frequency histogram for liver tumors across all datasets in AbdomenAtlas 3.0.** Each subplot represents a dataset with at least three tumor occurrences. The x-axis shows tumor size (cm), and the y-axis represents the number of tumors within each size range. The figure highlights the variability in tumor sizes annotated across datasets, and the significant presence of small tumors.

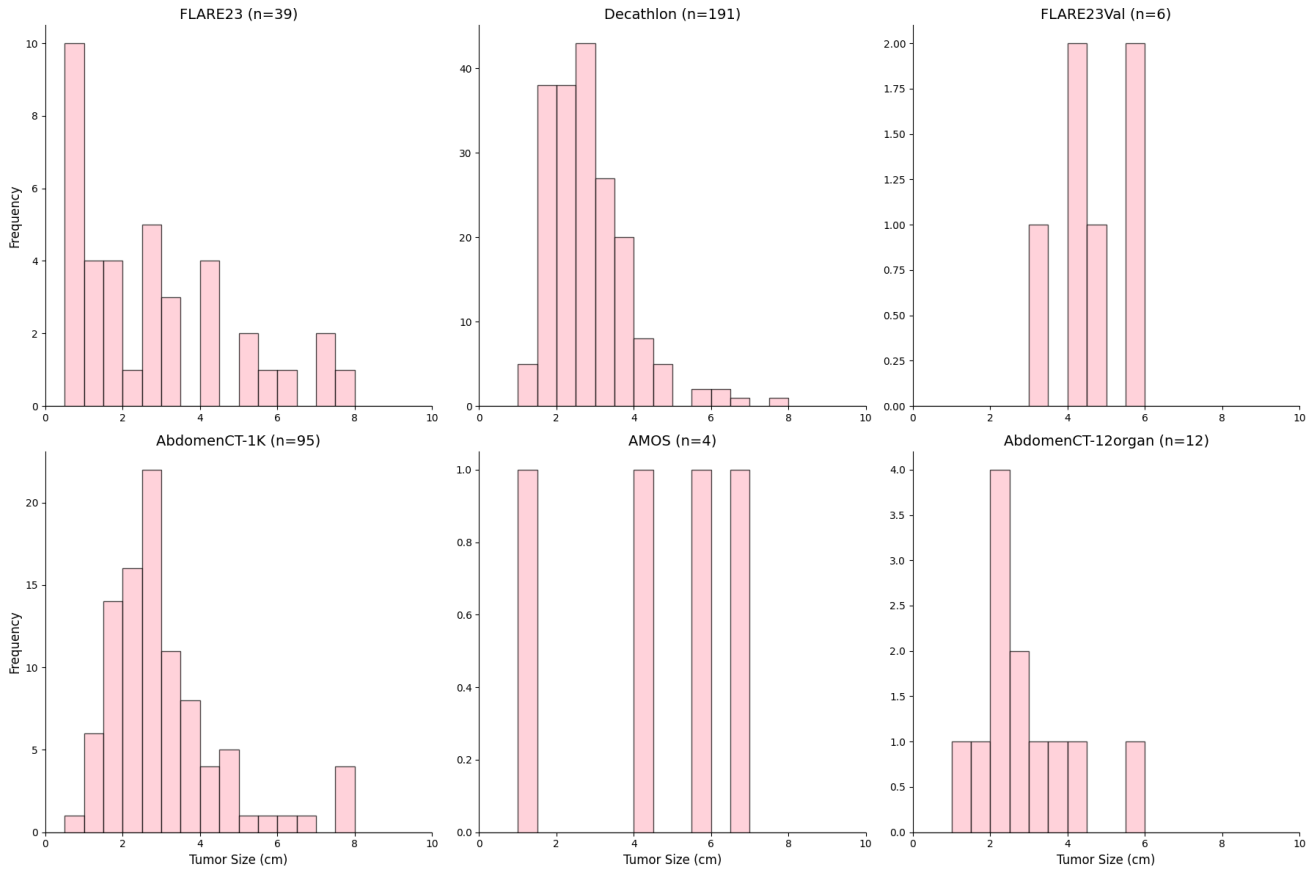

Figure 21. **Tumor size frequency histogram for pancreas tumors across all datasets in AbdomenAtlas 3.0.** Each subplot represents a dataset with at least three tumor occurrences. The x-axis shows tumor size (cm), and the y-axis represents the number of tumors within each size range. The figure highlights the variability in tumor sizes annotated across datasets, and the significant presence of small tumors.

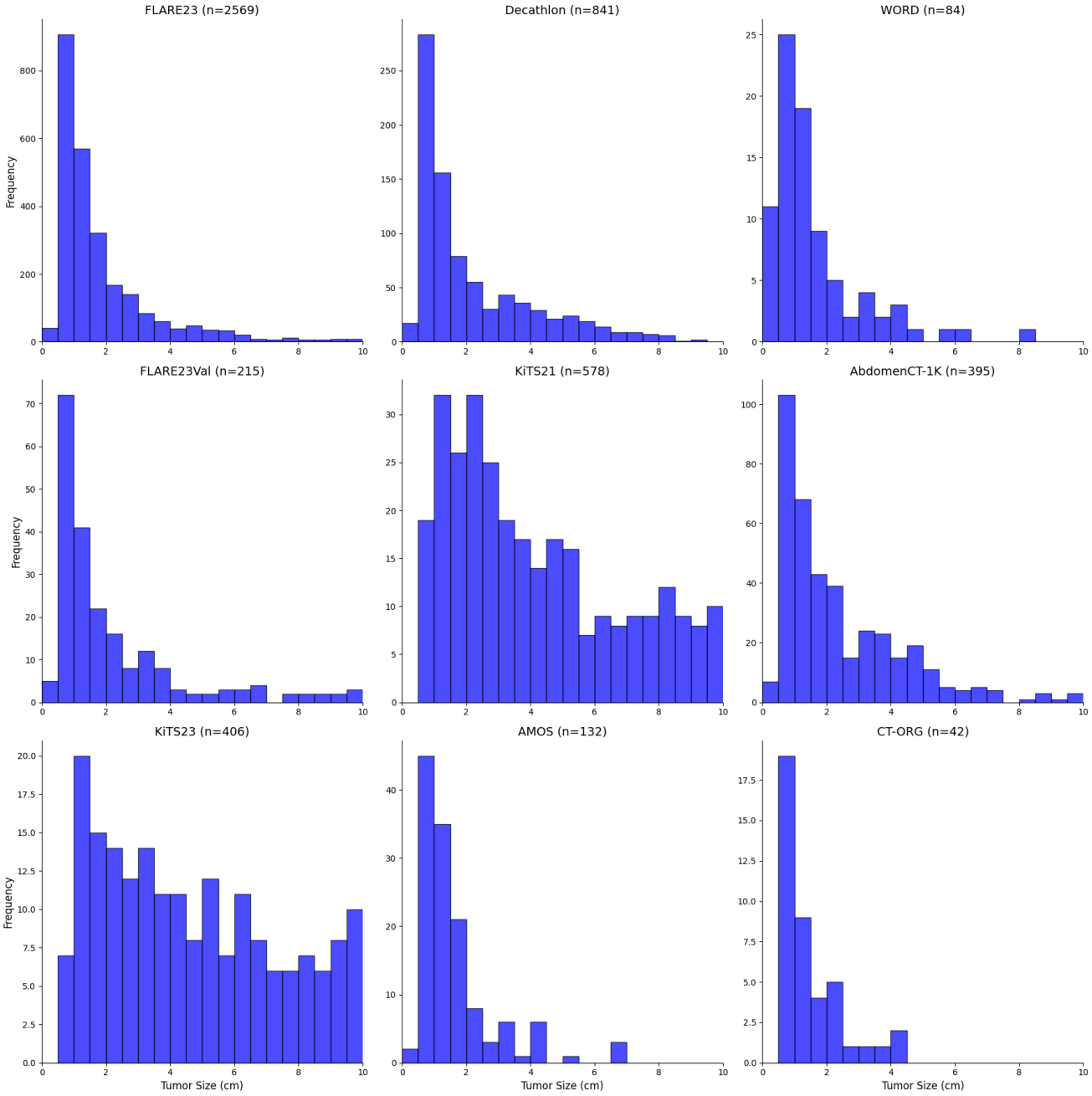

Figure 22. **Tumor size frequency histogram for kidney tumors across all datasets in AbdomenAtlas 3.0.** Each subplot represents a dataset with at least three tumor occurrences. The x-axis shows tumor size (cm), and the y-axis represents the number of tumors within each size range. The figure highlights the variability in tumor sizes annotated across datasets, and the significant presence of small tumors.