Medical Image Analysis 114 (2026) 104257

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Report Supervision ☆

Pedro R.A.S. Bassi [a,b,c], Wenxuan Li [a,b,c], Jakob Wasserthal [d], Jieneng Chen [a,e], Xinze Zhou [a], Zheren Zhu [f], Chuntung Zhuang [a], Sergio Decherchi [g], Andrea Cavalli [g,h,i], Kang Wang [f], Yang Yang [f], Alan Yuille [a], Zongwei Zhou [a,j,*]

[a] Department of Computer Science, Johns Hopkins University, United States of America
[b] Harvard Medical School, Harvard University, Boston, MA, United States of America
[c] Massachusetts General Hospital, Boston, MA, United States of America
[d] Clinic of Radiology and Nuclear Medicine, University Hospital Basel, Switzerland
[e] Department of Computer Science, Stanford University, United States of America
[f] Department of Radiology and Biomedical Imaging, University of California, San Francisco, United States of America
[g] Computational and Chemical Biology, Italian Institute of Technology, Italy
[h] Department of Pharmacy and Biotechnology, University of Bologna, Italy
[i] Centre Européen de Calcul Atomique et Moléculaire (CECAM), École Polytechnique Fédérale de Lausanne, Switzerland
[j] Department of Oncology, Johns Hopkins Medicine, United States of America

ARTICLE INFO

Dataset link: https://github.com/MrGiovanni/R-Super

Keywords:
Tumor detection
Tumor segmentation
Radiology reports
Weak supervision
Computed tomography

ABSTRACT

Segmentation models can surpass radiologists, classification models, and vision–language models in tumor detection. Furthermore, segmentation models outline tumors, allowing radiologists to better verify and trust the AI output. However, their main limitation is the scarcity of tumor masks: creating one 3D tumor mask takes up to 30 min, so most public CT datasets contain only a few hundred masks, and even the largest private datasets contain only a couple of thousand. Tumor masks are not produced in clinical routine, but radiology reports are. Public datasets contain tens of thousands of CT-Report pairs, and hospitals contain hundreds of thousands. These reports describe tumors in detail, providing large-scale, informative training data. Here, we introduce **Report Supervision (R-Super)**, a training framework that uses reports to directly supervise and improve tumor segmentation. R-Super introduces new loss functions that teach segmentation models to segment tumors that match report descriptions of tumor count, sizes, and locations. Reports are only used for training. We evaluated R-Super on kidney and pancreatic tumor segmentation, exploring diverse training data sizes, up to 41,418 CT-Report pairs plus 3488 CT-Mask pairs. On external validation, R-Super increased tumor detection F1-Score and segmentation DSC by up to +15% with respect to mask-only training. It also surpassed alternative methods such as CLIP and multi-task learning. It enabled pancreatic tumor detection on non-contrast CT without requiring tumor masks from non-contrast CT. In summary, by leveraging numerous readily available reports to supplement scarce masks, R-Super strongly improves AI performance when very few training masks are available (e.g., 50), and when many masks are available (e.g., 3488), unlocking scale in tumor segmentation. Project: https://github.com/MrGiovanni/R-Super.

1. Introduction

Tumor detection in CT remains difficult: over 50% of small pancreatic tumors are missed at first read, even by experienced radiologists (Hoogenboom et al., 2022). Kidney tumor detection is also challenging: studies have shown that, on non-contrast CT, radiologists missed 22% of small kidney tumors, while on contrast-enhanced CT (non-delayed), 11%–23% of tumors were missed (Gobara et al., 2019; Zeman et al., 1996). AI segmentation can support early detection. Segmentation models reveal where tumors are, allowing radiologists to better understand and verify the AI output. Additionally, segmentation models often surpass other AI models in tumor detection accuracy, including classification models (Hooper et al., 2023) and vision–language models (Bassi et al., 2025b). Finally, segmentation models can surpass human performance in tumor detection, since segmentation models can perceive subtle image abnormalities that may be imperceptible to humans (Cao et al., 2023).

☆ This article is part of a Special issue entitled: 'MICCAI 2025 Special Issue_invited only' published in Medical Image Analysis.

* Corresponding author at: Department of Computer Science, Johns Hopkins University, United States of America.
E-mail address: zzhou82@jh.edu (Z. Zhou).

https://doi.org/10.1016/j.media.2026.104257
Received 1 May 2026; Received in revised form 28 July 2026; Accepted 4 August 2026
Available online 20 August 2026
1361-8415/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Despite their large potential, tumor segmentation models face a long-standing limitation: the scarcity of tumor masks. Tumor segmentation masks are outlines of tumors in medical images, needed to train tumor segmentation models. Creating masks is expensive and time-consuming. Creating a single mask in a 3D scan can take a radiologist over 30 min. Creating a private dataset with 3488 pancreatic tumor masks (JHH dataset) required 8 radiologists, 5 years, and millions of dollars (Xia et al., 2022). Public datasets are even more limited. The majority of the public CT datasets have fewer than 100 tumor masks, few surpass 1000 tumor masks, and none exceeds 2000 masks per tumor type (Li et al., 2025b; Antonelli et al., 2021; Landman et al., 2017; Li et al., 2024; Chen et al., 2025). Tumor masks are scarce because radiologists do not create them as part of their job. However, radiologists routinely write radiology reports. Hospitals therefore hold hundreds of thousands of reports. The public Merlin dataset alone contains over 25,000 CT-Report pairs (Blankemeier et al., 2024). Crucially, reports contain information crucial for training segmentation models: tumor count, location, and size.

Reports are abundant and detailed, so they may be a key to unlocking scale in tumor segmentation, as text supervision has unlocked scale in general computer vision and natural language processing (Radford et al., 2021). Therefore, we ask: *can reports supplement masks and scale tumor segmentation datasets, improving tumor detection and segmentation?*

Previous work has used radiology reports to improve tumor segmentation. Studies extracted classification labels from reports (e.g., pancreatic tumor present vs absent). They used image-label pairs, along with image-mask pairs, to train multi-task learning (MTL) models for classification and segmentation (Zhang et al., 2021). Studies have also used reports to create foundational vision–language AI models, by pre-training them with CT-Report pairs and contrastive losses (e.g., CLIP) (Blankemeier et al., 2024). These models can be fine-tuned for tumor segmentation using tumor masks. However, MTL- and CLIP-based models share limitations: they do not use reports to directly improve tumor segmentation. Instead, reports optimize auxiliary tasks (classification or report-image contrastive alignment). The benefit of these auxiliary tasks to segmentation is indirect and sometimes small (Table 1).

Here, we introduce **R**eport **Super**vision (R-Super), a training framework that uses radiology reports to *directly optimize* tumor segmentation. R-Super trains segmentation models to segment tumors that match the tumor descriptions in reports. To this end, we introduce two loss functions, the Volume Loss and Ball Loss. These loss functions penalize differences between segmented tumors and reported tumor count, sizes, and locations (organ or organ sub-segments). This information is extracted from radiology reports by a large language model (LLM) before segmentation training. R-Super can train any segmentation architecture and uses reports only at training time, not at inference. Unlike previous methods, R-Super uses reports to directly optimize tumor segmentation. R-Super trains with both CT-Report pairs and CT-Mask pairs, together. For CT-Report pairs, R-Super trains with the Volume Loss and Ball Loss. For CT-Mask pairs, R-Super trains with standard segmentation losses (Dice and Cross-entropy). Therefore, with R-Super we can scale CT-Mask datasets with large amounts of CT-Report pairs. Training R-Super on these scaled datasets improves not only tumor segmentation but, importantly, also tumor detection.

To train R-Super, we built two new private CT-Report datasets, sourced from the University of California San Francisco (UCSF) Hospital and nearby institutions: UCSF-Train (6718 CT-Report pairs, kidney and pancreatic tumors) and UCSF-Huge (41,418 CT-Report pairs, 28,295 with pancreatic tumors). To test R-Super and to contribute data to the field, we built a new public CT-Report dataset (684 CT-Report pairs, 445 with pancreatic tumors).

We performed four main experiments. *First*, to demonstrate that reports improve tumor segmentation when few or many masks are available, we trained R-Super on UCSF-Train CT-Report pairs plus few (50), medium (344), and many (1.7K) CT-Mask pairs (Section 4.1). *Second*, to demonstrate that reports can scale a state-of-the-art public segmentation dataset, improving performance, we added 1848 public CT-Report pairs from Merlin (Blankemeier et al., 2024) to PanTS (Li et al., 2025b) (926 masks), the largest public pancreatic mask dataset (Section 4.2). *Third*, to demonstrate that reports can scale a massive private mask dataset, also improving performance, we added 41,418 CT-Report pairs from UCSF-Huge to JHH (Xia et al., 2022) (6212 masks), the largest private pancreatic mask dataset, annotated by 8 radiologists over 5 years (Section 4.3). *Fourth*, to demonstrate that R-Super enables tumor detection on non-contrast CT without non-contrast CT-Mask pairs, we trained it on contrast-enhanced CT-Mask pairs (JHH) plus CT-Report pairs containing both contrast-enhanced and non-contrast CTs (UCSF-Huge), and tested it on non-contrast CTs (Section 4.4). R-Super reduces the need for difficult manual tumor annotations on non-contrast CT. Experiments evaluated tumor detection (sensitivity, specificity) and segmentation (DSC, NSD), with external validation on hospitals unseen during training. R-Super consistently and substantially surpassed alternative training methods (CLIP (Radford et al., 2021), MTL (Zhang et al., 2021), Models Genesis (Zhou et al., 2021), standard mask-based segmentation (Gao et al., 2022), nnU-Net (Isensee et al., 2021), and report-based classification), some of which can also leverage reports. Our main contributions are:

1. R-Super introduces loss functions that optimize segmented tumors to be consistent with radiology reports. To the best of our knowledge, the R-Super losses are the first to use reports to directly optimize tumor segmentation in CT.
2. R-Super uses reports to improve tumor detection and segmentation whether few or many masks are available for training (up to +15% F1-Score and DSC over training with masks alone, Table 2).
3. R-Super uses reports to improve performance even on top of state-of-the-art mask datasets, both public (PanTS, Section 4.2) and private (JHH, Section 4.3).
4. By supplementing contrast-enhanced CT-Mask datasets with non-contrast CT-Report pairs, R-Super enables the detection of pancreatic tumors on non-contrast CT scans (Section 4.4).
5. We release a new public CT-Report dataset, with 684 CT-Report pairs, focusing on pancreatic tumors (Section 4.3).

2. Related work

We compare R-Super against six alternative training methods for tumor segmentation: standard mask-supervised segmentation (MedFormer Gao et al., 2022, nnU-Net Isensee et al., 2021), self-supervised pre-training (Models Genesis Zhou et al., 2021), CLIP with segmentation fine-tuning (CLIP Radford et al., 2021), multi-task learning (Chen et al., 2019), and report-guided pseudo-labels (Bosma et al., 2023). Most of these alternative methods either ignore reports or use them to supervise an auxiliary training task, whereas R-Super uses reports to directly optimize the tumor segmentation output. We develop these comparisons below. We then explain how this paper extends our previous work.

2.1. Alternative training methods

CLIP. A common way to use reports for tumor segmentation today is contrastive vision–language pre-training (CLIP). Most medical vision–language models (VLMs) are pre-trained with variants of the CLIP loss (Blankemeier et al., 2024; Hamamci et al., 2024; Sellergren et al., 2025). These losses pre-train the AI model by giving it a set of images and a set of texts and asking it to find the correct image-text pairs across the two sets. After pre-training, the models can be fine-tuned for tumor segmentation using masks. The features learned by the model during pre-training may help it learn segmentation, but this benefit is indirect and often small (Table 1). A low CLIP loss during pre-training does not guarantee high-quality segmentation. In contrast, the Report

Supervision losses supervise the segmentation output itself, penalizing the model whenever the segmented tumors disagree with the tumor descriptions in the report. A low R-Super loss therefore corresponds, by construction, to segmentations that match the reports.

Multi-task Learning. Multi-task learning (MTL) frameworks (Chen et al., 2019) proposed that classification can improve segmentation. In MTL, both segmentation and classification are learned together, by the same model. Features that the model learns for classification can also help segmentation. However, there is no guarantee that one task will help the other. In some cases, this mutual help was marginal or absent (Bassi et al., 2024a). Like CLIP, MTL uses reports to indirectly improve segmentation, so a model can achieve strong classification and poor segmentation. In contrast, R-Super directly uses reports to supervise segmentation.

Weakly supervised learning and promptable segmentation models. Medical segmentation methods have reduced segmentation mask requirements using alternative, weaker sources of supervision, such as bounding boxes and scribbles. DeepCut learns segmentation from bounding boxes (Rajchl et al., 2017), while scribble-supervised methods learn from sparse lines drawn within target structures (Can et al., 2018). More recently, promptable foundation models such as SAM and MedSAM convert points or bounding boxes into segmentation masks (Ma et al., 2024). These approaches can reduce annotation effort relative to per-voxel segmentation masks, but their spatial annotations or prompts must still be deliberately created and are not routinely available in hospital archives. In contrast, radiology reports are routinely produced and available at scale in hospitals. R-Super uses reports to supervise tumor segmentation without requiring new annotations. Some weak-supervision methods use their annotations (e.g., bounding boxes) only during training, whereas promptable models such as SAM and MedSAM require a point or bounding box at inference to identify the object to segment (e.g., a tumor). This inference-time requirement is particularly limiting for tumor detection: SAM assumes that a radiologist or another AI system has already detected and roughly localized the tumor and is therefore not directly designed to assist in tumor detection.

Self-supervised Learning. Self-supervised training methods can learn from CT scans alone, using neither masks nor reports. Masked autoencoders (e.g., Models Genesis Zhou et al., 2019, 2021; Haghighi et al., 2020) receive images with parts masked out and learn to reconstruct the unmasked images, thus learning structural and geometric relationships in medical images. Other self-supervised methods use different pre-training objectives. Rubik's Cube+ trains a model to recover the order and orientation of shuffled 3D image cubes and identify masked cubes (Zhu et al., 2020). Chaitanya et al. (2020) use global and local contrastive learning. Swin UNETR combines masked-volume inpainting, rotation prediction, and contrastive learning (Tang et al., 2022). Here, we adopt Models Genesis as a self-supervised baseline because it is a well-established 3D medical self-supervised method that has been independently validated on multiple downstream tasks (Wald et al., 2024), making it a strong representative of image-only self-supervised pre-training. After pre-training, self-supervised models can be fine-tuned for tumor segmentation using masks. However, this approach ignores reports, which provide detailed tumor information. In contrast, R-Super is designed to exploit this detailed information, specifically the tumor count, sizes, and locations.

Report-guided pseudo-labels (Bosma et al., 2023) trains a segmentation model on images with tumor masks, and uses this model (teacher) to create pseudo-masks for images without masks. Then, these pseudo-masks are filtered according to reports: only the top-n most confident tumors are kept in each mask, where n is the number of tumors in the report. The filtered pseudo-masks (plus the original masks) are used to train a second segmentation model (student). This filtering scheme reduces false positives in the pseudo-masks, but does not address false negatives, i.e., tumors that the teacher model missed. Thus, masks with false negatives were discarded from the student's training set (Bosma et al., 2023). However, removing these cases removes precisely the hard-to-detect tumors that would be most informative to learn from. In contrast, R-Super can learn from these hard-to-detect cases, as we designed the Volume Loss to strongly penalize false negatives (Section 3.2). Report-guided pseudo-labels also do not use important report details, such as tumor sizes and locations. In contrast, R-Super uses these details in its loss functions (Section 3.3).

2.2. Our previous work

Our MICCAI conference paper (Bassi et al., 2025a) first introduced the Report Supervision framework, proposing the Volume Loss and the Ball Loss. This paper extends the preliminary work substantially with the following improvements:

1. *New public dataset:* 684 CT-Report pairs, including 445 pancreatic tumor cases.
2. *New experiment with public data:* we show that adding CT-Report pairs, from the *public* Merlin dataset, to the largest *public* pancreatic tumor segmentation dataset (PanTS Li et al., 2025b) improves tumor detection and segmentation.
3. *New large-scale experiment:* we show that adding 41,418 CT-Report pairs to the largest *private* pancreatic tumor mask dataset (6212 CT-Mask pairs, created by 8 radiologists in 5 years) further improves AI performance.
4. *Experiments on non-contrast CT:* we demonstrate that adding non-contrast CT-Report pairs to existing contrast-enhanced CT-Mask datasets allows R-Super to detect pancreatic cancer in non-contrast CT.
5. *Formalized methodology.* We provide a more complete formulation of the loss functions, with additional equations and a clearer description of the framework.

3. R-super

R-Super is a training framework that enforces segmented tumors to match the tumor descriptions in radiology reports. It is summarized in Fig. 1. First, an LLM extracts the descriptions of tumor diameters, locations (e.g., pancreatic head), and counts from radiology reports (Section 3.1). Second, this information is used as ground-truth for two loss functions, the Volume Loss and the Ball Loss, which train the tumor segmentation model (any architecture, we used MedFormer Gao et al., 2022). The *Volume Loss* (Section 3.2) penalizes differences between the tumor volume estimated from the model's segmentation output and the volume estimated from the report. The *Ball Loss* (Section 3.3) penalizes differences in tumor location (e.g., pancreatic head, body, or tail), count, and diameter.

The Ball Loss is stricter than the Volume Loss, because it requires segmented tumors to match multiple tumor characteristics from the reports. In contrast, the Volume Loss is easier to optimize. Therefore, we apply the Volume Loss as deep supervision (e.g., applying it at the second layer of the MedFormer decoder), and we apply the Ball Loss at the final layer of the segmentation model. This design enables easy optimization while ensuring that the final segmentation output matches the detailed tumor descriptions in radiology reports. We train R-Super with CT-Mask pairs plus numerous CT-Report pairs. For CT-Report pairs, we use the Ball Loss and Volume loss; for CT-Mask pairs, we use standard segmentation losses, Dice and cross-entropy.

The Volume Loss and Ball Loss are guided not only by reports but also by organ and organ sub-segment masks. These masks localize in the image the diseased organs mentioned in the report, allowing our losses to teach the tumor segmentation model to segment tumors only within these diseased organs. We generate the organ/sub-segment masks before training the tumor segmentation model. To this end, we use AtlasNet, an nnU-Net trained on AbdomenAtlas 3 (9262 CT scans)

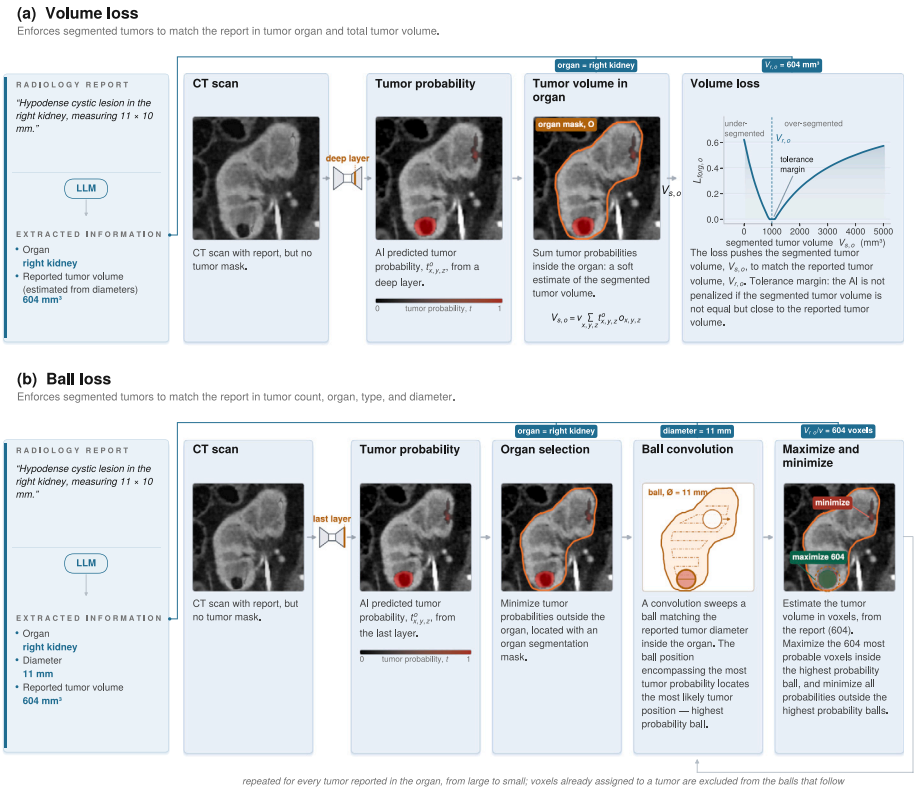

Fig. 1. R-Super trains tumor segmentation models using information from radiology reports, with Volume Loss and Ball Loss guiding segmented tumors to match reported tumor size, location, and number. (a) *The Volume Loss (deep supervision) aligns segmented tumors with volumes and locations from reports.* It consists of three steps. **(I)** A large language model (LLM) extracts tumor location (organ/sub-segment) and diameters from the report. From these diameters, we estimate the total tumor volume in each organ o as $V_{r,o}$, using ellipsoid approximations. **(II)** From a deep layer of the segmenter, we compute tumor probabilities via a $1 \times 1 \times 1$ convolution and sigmoid/softmax. We sum the tumor probabilities inside the organ (located with pre-saved organ masks) to estimate the segmented tumor volume, $V_{s,o}$. **(III)** The Volume Loss penalizes differences between the segmented and reported tumor volumes ($V_{s,o}$ and $V_{r,o}$). As shown in the right panel (for $V_{r,o} = 1000$), the loss is zero when $V_{s,o}$ is close to $V_{r,o}$, allowing for some tolerance due to uncertainty in report-based volume estimates. **(b)** *The Ball Loss enforces segmented tumors to match detailed tumor information from reports—tumor count, locations, diameters, and volumes.* It consists of four steps. **(I)** A large language model (LLM) extracts this tumor information from the report. **(II)** A Ball Convolution — a simple convolution with a fixed spherical kernel matching the reported tumor diameter — slides over the segmenter output to find the highest-probability tumor region. **(III)** Within this region, the loss maximizes probabilities in the top-N most probable voxels, where N is the tumor volume estimated from the report (in voxels). These voxels can form any shape that stays within the ball. **(IV)** The process (II–III) is repeated for all reported tumors, largest to smallest. Already-assigned voxels are ignored in later iterations to prevent overlap. Finally, the loss minimizes tumor probabilities in unassigned voxels (background). It focuses more on central tumor voxels and avoids penalizing uncertain tumor borders.

(Bassi et al., 2025b) to segment organs such as the pancreas and its sub-segments (pancreas head/body/tail). The kidney also has sub-segments, including the renal cortex, medulla, and pelvis. However, our nnU-Net does not segment these sub-segments separately. Therefore, for the kidneys, we use organ masks (which separate the right and left kidneys). Organ masks can reliably guide our tumor segmentation losses because organ segmentation is an easier task, and public AI models segment organs with much higher accuracy than tumors (Qu et al., 2023). For example, nnU-Net trained on AbdomenAtlas achieves 82.9% DSC for pancreas segmentation (Bassi et al., 2024b). Still, to compensate for inaccuracies in organ/sub-segment masks and to compensate for tumors that grow beyond organ boundaries, we dilate the organ/sub-segment masks (by 2 cm). Even in ablation studies with organ masks generated by an nnU-Net trained on only 50 CT scans, R-Super maintained high tumor detection and segmentation performance (Table 5). Although organ masks improve the R-Super performance by making the Volume Loss and Ball Loss more precise, these losses remain effective even without organ masks (Table 5).

3.1. LLM for extracting tumor information

We use a large language model (LLM) to extract tumor information from reports, because LLMs can understand both context and semantics. This understanding is critical because reports vary widely across hospitals, locations, time periods, and radiologist preferences. These variations affect both style and wording: some reports are highly structured, while others are narrative, and the same finding can be expressed with different terms (e.g., a "malignant mass" or a "lesion consistent with the patient's known pancreatic cancer"). Semantic understanding allows the LLM to recognize that these different expressions refer to the same underlying finding. Context is also essential, as tumor measurements may refer to current tumors or to tumors in prior exams. By combining semantic and contextual understanding, and by following clinical guidelines specified in their prompts, LLMs can accurately extract tumor information from reports.

As the LLM, we use Llama 3.1 70B AWQ (Dubey et al., 2024). Its prompts explain how to interpret reports, and they were developed through iterative error analysis and refinement, in a collaboration between computer scientists and radiologists. Prompts are available in Appendix D. To reduce computational cost, the LLM is used only once, before training the segmenter. Our prompts instruct the LLM to return structured outputs, where the extracted tumor information fills a template. This template is then automatically parsed into a table, which will be used as ground-truth for the Volume Loss and Ball Loss. In addition to structured outputs, the LLM returns justifications for their answers and supporting quotes from the reports.

We use in-context learning, providing the LLM with examples of reports and their desired outputs. By not fine-tuning the LLM, we avoid

overfitting it to the style of reports in our datasets, enabling easy generalization across hospitals, and keeping computational costs low. A radiologist evaluated this LLM on 447 reports (182 with tumors) and found that it achieves 96% accuracy in extracting tumor information (Bassi et al., 2025b). The tumor detection and segmentation performance of R-Super shows minimal degradation even with LLM error rates over 10% (Table 5).

3.2. Volume Loss

The Volume Loss minimizes the dissimilarity between the volume of segmented tumors ($V_{s,o}$) and the tumor volume estimated from radiology reports ($V_{r,o}$). Specifically, $V_{r,o}$ is the combined volume of all tumors reported in the organ or sub-segment o (e.g., pancreatic head), and $V_{s,o}$ is the combined volume of all tumors segmented in organ or sub-segment o. By minimizing the dissimilarity between $V_{s,o}$ and $V_{r,o}$, the Volume Loss teaches the segmentation model to match the volume of tumors mentioned in the report, to avoid false positives (segmenting a tumor when the report mentions no tumor), and to avoid false negatives (segmenting no tumor when the report mentions tumors). The Volume Loss also penalizes tumors segmented outside of the organ/sub-segment o.

Main Equation. Eq. (1) defines the Volume Loss for an organ o (or organ sub-segment, we will say organ for simplicity). When the report mentions tumors in organ o, the Volume Loss has two terms: $L_{\text{forg},o}(V_{s,o}, V_{r,o})$, which penalizes the dissimilarity between segmented and reported tumor volumes, and $L_{\text{bkg},o}(\mathbf{T}^o)$, which penalizes tumors segmented outside organ o. When no tumors are reported in organ o, the loss penalizes any segmented tumor, using a per-voxel cross-entropy with target zero, $\text{CE}(\mathbf{T}^o, \mathbf{0})$. Here, $\mathbf{T}^o$ is the segmentation model's output channel for tumors in organ o (e.g., the pancreatic tumor channel), made of per-voxel estimated tumor probabilities. When applying the Volume Loss as deep supervision, $\mathbf{T}^o$ is produced by a 1×1×1 convolution with sigmoid activation, applied to an intermediate layer of the segmentation model's decoder, then up-sampled to the size of the segmentation model's input. The total Volume Loss is the sum of $L_{\text{vol},o}$ over all organs of interest $\mathcal{O}$ (Eq. (2)). Next, we break down the calculation of all terms in the Volume Loss.

$$L_{\text{vol},o} = \begin{cases} L_{\text{forg},o}(V_{s,o}, V_{r,o}) \\ \quad + L_{\text{bkg},o}(\mathbf{T}^o) & \text{if tumors reported in } o \\ \text{CE}(\mathbf{T}^o, \mathbf{0}) & \text{otherwise} \end{cases} \tag{1}$$

$$L_{\text{vol}} = \sum_{o \in \mathcal{O}} L_{\text{vol},o} \tag{2}$$

Estimating tumor volumes from radiology reports ($V_{r,o}$). Most radiology reports do not report tumor volumes, but they often report tumor diameters. One, two, or three perpendicular diameters are reported per tumor (d_1, d_2, d_3). For instance, the WHO standard (Miller et al., 1981) uses two diameters, and RECIST uses one (Eisenhauer et al., 2009). When the report provides a single diameter for the tumor, we estimate its volume as a ball's volume, $d_1^3\pi/6$. When 3 diameters are provided, we use the ellipsoid volume, $d_1 d_2 d_3 \pi/6$. When 2 diameters are provided, we estimate the third as the average of the other two, and use the ellipsoid volume formula again, $d_1 d_2((d_1 + d_2)/2)\pi/6$. We calculate the volume for each tumor reported in organ o, and sum them, obtaining the total reported tumor volume, $V_{r,o}$.

Estimating tumor volumes from segmentation ($V_{s,o}$). To estimate the total volume of all tumors segmented inside organ o, $V_{s,o}$, we could simply count the tumor voxels segmented inside organ o and multiply the result by the volume of one voxel, v. This is similar to element-wise multiplying the tumor segmentation output ($\mathbf{T}^o = [t^o_{x,y,z}]$) with the organ segmentation mask ($\mathbf{O} = [o_{x,y,z}]$), summing over the spatial dimensions (x, y, z), and multiplying the result by the voxel volume v (Eq. (3)). Here, $\mathbf{O}$ is a binary mask, equal to 1 inside organ o and 0 outside. Multiplying by $\mathbf{O}$ ensures that $V_{s,o}$ counts only tumors segmented inside organ o. If $\mathbf{T}^o$ were binary, Eq. (3) would be exactly equivalent to counting tumor voxels. Since $\mathbf{T}^o$ contains tumor probabilities, Eq. (3) provides a soft estimate of tumor volume. Importantly, this estimate is differentiable, and a correct tumor segmentation would yield $V_{s,o} = V_{r,o}$.

$$V_{s,o} = v \sum_{x,y,z} t^o_{x,y,z} o_{x,y,z} \tag{3}$$

Tumor volume dissimilarity ($L_{\text{forg},o}(V_{s,o}, V_{r,o})$). To enforce the volume of segmented tumors to match the tumor volume estimated from the radiology reports, we minimize the dissimilarity between $V_{s,o}$ and $V_{r,o}$. Specifically, we minimize the dissimilarity function $L_{\text{forg},o}(V_{s,o}, V_{r,o})$ defined in Eq. (4). In the equation, E is a small constant (set to 500 mm^3) that provides numerical stability for small $V_{r,o}$, and $\tau \in (0, 1)$ is a tolerance margin (we set $\tau = 0.1$). Three key properties of $L_{\text{forg},o}(V_{s,o}, V_{r,o})$ are observable in its plot in Fig. 1. *First*, $L_{\text{forg},o}(V_{s,o}, V_{r,o})$ has a steep but finite gradient at $V_{s,o} = 0$ (if $V_{r,o} \neq 0$). Thus, the Volume Loss strongly penalizes false negatives (missed tumors) while remaining numerically stable. *Second*, the gradient is less steep when $V_{s,o} > V_{r,o}$, because strong gradients in this region could enforce false negatives, by pushing the segmentation model towards $V_{s,o} = 0$. *Third*, the Volume Loss has a tolerance margin: $L_{\text{forg},o}(V_{s,o}, V_{r,o})$ and its gradient are zero when $V_{s,o}$ is very close to $V_{r,o}$ (we use a tolerance, τ, of 10%). This margin compensates for inaccuracies in $V_{r,o}$, which arise from human error in measuring tumor diameters in reports, and from the approximation errors when we convert those diameters into volumes. Therefore, the segmentation model is not penalized if the tumors it segments have a volume that is close but not equal to $V_{r,o}$.

$$\begin{aligned} L_{\text{forg},o}(V_{s,o}, V_{r,o}) = \max\{ & L'_{\text{forg},o}(V_{s,o}, V_{r,o}) \\ & - L'_{\text{forg},o}((1-\tau)V_{r,o}, V_{r,o}), 0\} \end{aligned} \tag{4}$$

$$L'_{\text{forg},o}(V_{s,o}, V_{r,o}) = \frac{|V_{s,o} - V_{r,o}|}{V_{s,o} + V_{r,o} + E} \tag{5}$$

Avoiding tumors in wrong locations ($L_{\text{bkg},o}(\mathbf{T}^o)$). The minimization of $L_{\text{forg},o}(V_{s,o}, V_{r,o})$ teaches the segmentation model to segment tumors inside the organ o where the report mentions tumors. However, we must also penalize tumors segmented outside organ o (e.g., pancreatic tumors segmented outside of the pancreas). To this end, the Volume Loss includes the term $L_{\text{bkg},o}(\mathbf{T}^o)$, defined in Eq. (6). This term uses voxel-wise cross-entropy to penalize tumors segmented outside the organ o, located by its organ segmentation mask $\mathbf{O}^o$. Specifically, we element-wise multiply $(1 - \mathbf{O}^o) \odot \mathbf{T}^o$, capturing tumors segmented outside organ o. Afterwards, we penalize these tumors with a cross-entropy against a zero target. Organ masks were dilated (by 2 cm) to compensate for tumors that grow beyond organ borders.

$$L_{\text{bkg},o}(\mathbf{T}^o) = \text{CE}((1 - \mathbf{O}) \odot \mathbf{T}^o, \mathbf{0}) \tag{6}$$

Reports missing tumor count or diameters. The estimation of $V_{r,o}$ requires the report to provide diameters for every tumor in organ o. When a report mentions tumors in o but omits the number of tumors or their diameters, we fall back to a relaxed form of the loss: it requires the segmentation model to segment *at least one* small tumor inside organ o, but does not penalize segmenting a larger tumor or multiple tumors. The full reformulation is given in Appendix A.

Input cropping. Due to GPU memory constraints, tumor segmentation models are often trained on patches rather than full 3D images. This creates a problem for the Volume Loss (and for the Ball Loss): if organ o is only partially contained in a training patch, we cannot know whether a tumor reported in o is inside the patch or outside. The Volume Loss and the Ball Loss cannot be applied in this case. To avoid this, we use the organ masks to guide cropping, ensuring that each training patch fully contains at least one organ o. The cropped organ o is randomly selected, with high probability (80%) to crop on organs with tumors. If the patch partially contains a second organ with tumors, we skip the Volume Loss and the Ball Loss for this second organ.

3.3. Ball Loss

The Ball Loss encourages the segmentation model to segment tumors that match radiology reports in tumor count, location (organ/organ sub-segment), and diameters. The Ball Loss localizes, in the output of the tumor segmentation model, the most likely location for each tumor mentioned in the radiology report. Then, each of these located tumors is optimized to match the corresponding tumor diameter in the report. The Ball Loss can be defined by a sequence of procedures, illustrated in Fig. 1 and described below.

Locate the organ o and penalize tumors segmented outside of it. We first locate the organ (or organ sub-segment) o in the output of the tumor segmentation model, using the pre-saved organ segmentation mask $\mathbf{O}$. In case the report mentions no tumor in organ o, we use cross-entropy with zero target to minimize the entire tumor segmentation output for that organ, $L_{\text{ball}} = \text{CE}(\mathbf{T}^o, \mathbf{0})$. Otherwise, we multiply the tumor segmentation output ($\mathbf{T}^o$, per-voxel probabilities for tumors in organ o) by the organ segmentation mask for organ o, $\mathbf{O} \odot \mathbf{T}^o$. This step removes tumors segmented outside organ o before the next step, when we locate the tumors reported in organ o.

Find the most likely location for a tumor i described in the report (Ball Convolution). We begin by searching for the largest tumor that the report describes in organ o, tumor $i = 0$ with diameter d_i (largest diameter reported for tumor i). To find the tumor's most likely location, we use a convolution that we call *Ball Convolution*. It is a standard 3D convolution (stride 1, padding 1, odd sized kernel), but its kernel is not learnable. Its kernel $\mathbf{k}_i$ is a binary ball with diameter matching the tumor diameter d_i in the report (Eq. (7)). The convolution operation moves this ball over the output of the tumor segmentation model, $\mathbf{T}^o$ (which we previously multiplied by the organ mask, $\mathbf{O}$), as shown in Eq. (8) (where $*$ indicates convolution). At each location, the ball sums the per-voxel tumor probabilities inside it. Therefore, when the ball is at the most likely location for a tumor of diameter d_i (the reported tumor diameter), the output of the Ball Convolution is the highest. I.e., the maximum output of the Ball Convolution, $\mathbf{c}_i = [c_x, c_y, c_z]$ (Eq. (9)), defines the most likely center for a ball encompassing the tumor of the reported diameter d_i. This center must be inside the organ o, due to the multiplication of the tumor segmentation output $\mathbf{T}^o$ and the organ mask $\mathbf{O}$ in the last step. After finding $\mathbf{c}_i$, we define the *highest probability ball*, $\mathbf{H}_i$, as a ball of diameter $(1+\mu)d_i$ and center $\mathbf{c}_i$ (Eq. (10)). We expand the highest probability ball by $\mu = 20\%$ to account for inaccuracies in the reported diameter d_i. Even if a tumor has an irregular shape, it is expected to fit inside a ball of diameter $(1+\mu)d_i$. That is because d_i is defined as the maximum distance between two tumor points in any axial plane, according to most radiology reporting guidelines (e.g., WHO and RECIST). Our $(1+\mu)d_i$ expansion accommodates diameter measurement inaccuracies and larger out-of-plane tumor extents.

$$\mathbf{k}_i(u,v,w) = \begin{cases} 1, & \text{if } \sqrt{u^2+v^2+w^2} \le \dfrac{d_i}{2} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

$$\mathbf{B}_i = (\mathbf{O} \odot \mathbf{T}^o) * \mathbf{k}_i \tag{8}$$

$$\mathbf{c}_i = [c_x, c_y, c_z] = \underset{(x,y,z)}{\text{argmax}}\, \mathbf{B}_i(x,y,z) \tag{9}$$

$$\mathbf{H}_i(x,y,z) = \begin{cases} 1, & \text{if } \|(x,y,z) - \mathbf{c}_i\| \le \dfrac{d_i(1+\mu)}{2} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Locate the N_i most likely voxels for a tumor i described in the report. The Ball Loss does not maximize all voxels inside the highest probability ball $\mathbf{H}_i$, because we do not want to teach the segmentation model to segment perfectly spherical tumors. Instead, the Ball Loss maximizes the top-N_i voxels with the highest predicted tumor probability inside the highest probability ball. These top-N_i voxels can assume any shape that *fits inside* the highest probability ball. We obtain N_i by converting the volume of tumor i, estimated from the report (Section 3.2), to a voxel count, and reducing it by a small margin (20%). From these top-N_i voxels we build a binary mask $\mathbf{M}_i$, with value 1 at the top-N_i voxels and 0 elsewhere (Eq. (12)). We then build a second mask $\mathbf{M}'_i$, with value 1 at the $[N_{i+1}, \nu n]$ (we set $\nu = 1.2$) voxels with highest tumor probability inside the highest probability ball (Eq. (13)). $\mathbf{M}_i$ thus marks the most probable tumor voxels for tumor i, and $\mathbf{M}'_i$ marks a tolerance margin around $\mathbf{M}_i$. This tolerance margin accounts for uncertain tumor borders and tumor measurement error in the reports.

$$\widetilde{\mathbf{T}}^o_i = \mathbf{T}^o \odot \mathbf{H}_i \tag{11}$$

$$\mathbf{M}_i(x,y,z) = \begin{cases} 1, & \text{if } (x,y,z) \in \text{top}_{[1,\,N_i]}(\widetilde{\mathbf{T}}^o_i) \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

$$\mathbf{M}'_i(x,y,z) = \begin{cases} 1, & \text{if } (x,y,z) \in \text{top}_{[N_i+1,\,\nu N_i]}(\widetilde{\mathbf{T}}^o_i) \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

Maximize and minimize tumor probabilities. In case the report mentions a single tumor i in organ o, the Ball Loss maximizes the segmentation model's output $\mathbf{T}^o$ at the N_i most likely tumor voxels (mask $\mathbf{M}_i$), leaves the surrounding margin (mask $\mathbf{M}'_i$) unpenalized, and minimizes $\mathbf{T}^o$ everywhere else. We achieve all three behaviors with a single cross-entropy plus Dice loss, using $\mathbf{M}_i$ as the target and $\mathbf{T}^o$ as the prediction (Eq. (14), assuming $\mathbf{M} = \mathbf{M}_i$ and $\mathbf{M}' = \mathbf{M}'_i$ in case of a single tumor in organ i). To avoid penalizing the margin voxels $\mathbf{M}'_i$, we multiply $\mathbf{T}^o$ by $(\mathbf{1} - \mathbf{M}'_i)$ before computing the loss, zeroing the gradient at those voxels (Eq. (14)). The resulting loss maximizes tumor probabilities at the most likely tumor voxels given the report and the segmentation model's output, accounting for uncertainty.

$$\mathcal{L}'_{\text{ball},o} = \text{CE}((1-\mathbf{M}') \odot \mathbf{T}^o, \mathbf{M}) + \text{Dice}((1-\mathbf{M}') \odot \mathbf{T}^o, \mathbf{M}) \tag{14}$$

$$\mathcal{L}_{\text{ball},o} = \begin{cases} \mathcal{L}'_{\text{ball},o}, & \text{if tumors are reported in } o \\ \text{CE}(\mathbf{T}^o, \mathbf{0}), & \text{otherwise} \end{cases} \tag{15}$$

$$L_{\text{ball}} = \sum_{o \in \mathcal{O}} L_{\text{ball},o} \tag{16}$$

Reports with multiple tumors. When a report mentions multiple tumors in organ o, we locate them sequentially from largest to smallest, applying the procedure above for each tumor i: Ball Convolution, create masks $\mathbf{M}_i$ and $\mathbf{M}'_i$. To avoid mask overlaps between tumors, after locating tumor i, we zero its most probable voxels and uncertainty border ($\mathbf{M}_i \bigcup \mathbf{M}'_i$) in the segmentation output, $\mathbf{T}^o \leftarrow \mathbf{T}^o \odot (\mathbf{1} - (\mathbf{M}_i \bigcup \mathbf{M}'_i))$. Thus, the next Ball Convolution (to locate tumor $i+1$) cannot place its center inside the voxels already assigned to tumor i. Once all reported tumors in o are located, we take the union of their per-tumor masks to form the combined masks $\mathbf{M}$ and $\mathbf{M}'$ (Eq. (17)). Then, these combined masks guide the cross-entropy and dice losses in Eq. (14). When tumors are reported across multiple organs, the final Ball Loss is the sum of the per-organ losses (Eq. (16)).

$$\mathbf{M} = \bigcup_{i=1}^{K^o} \mathbf{M}_i, \qquad \mathbf{M}' = \bigcup_{i=1}^{K^o} \mathbf{M}'_i \tag{17}$$

Reports missing tumor count or diameters. The presented formulation of the Ball Loss needs the report to provide diameters for all tumors in the organ o. When a report misses tumor diameters or counts, we use a high-tolerance variant of the Ball Loss (Appendix A). It encourages the segmentation model to segment at least one small tumor in o. However, it does not penalize the segmentation model for segmenting larger tumors nor many tumors.

As final details, to further compensate for uncertainty at tumor borders, our implementation of cross-entropy (Eq. (14)) is weighted, giving higher weights to voxels where the segmentation model predicts higher tumor probabilities. The Ball Convolution kernel also has a small Gaussian decay, with values larger near the ball center, slightly smaller near the ball boundary, and zero outside the ball. Because predicted tumor probabilities tend to peak at tumor centers, this decay improves

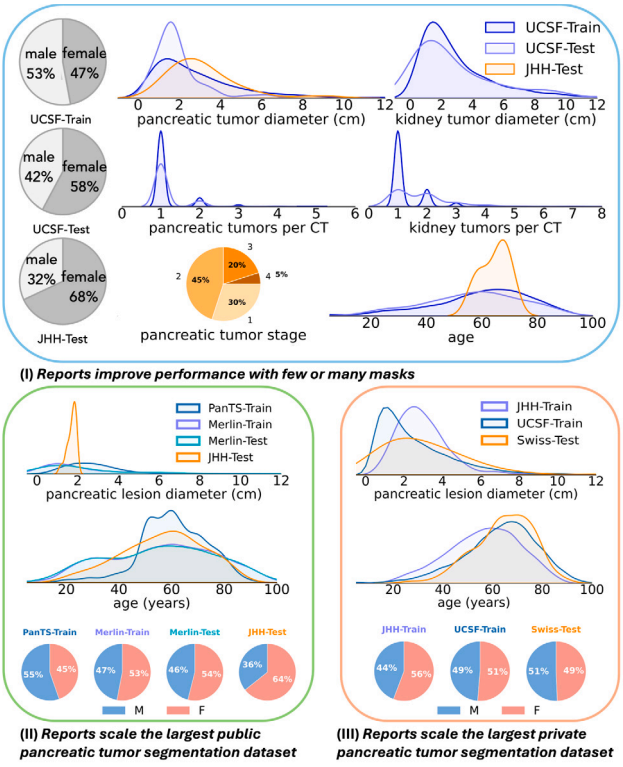

Fig. 2. Dataset overview (Sections 3.1–3.3). Summary of all datasets, including CT–mask and CT–report cohorts from multiple institutions. In (I), (II) and (III), we show the number of tumor scans, tumor types (pancreatic and kidney), and distributions of tumor size, tumor count per scan, and patient demographics (age and sex) across training and test sets (UCSF, JHH, and public datasets). Radiology reports enable large-scale dataset construction (Li et al., 2025a), which R-Super leverages for segmentation training.

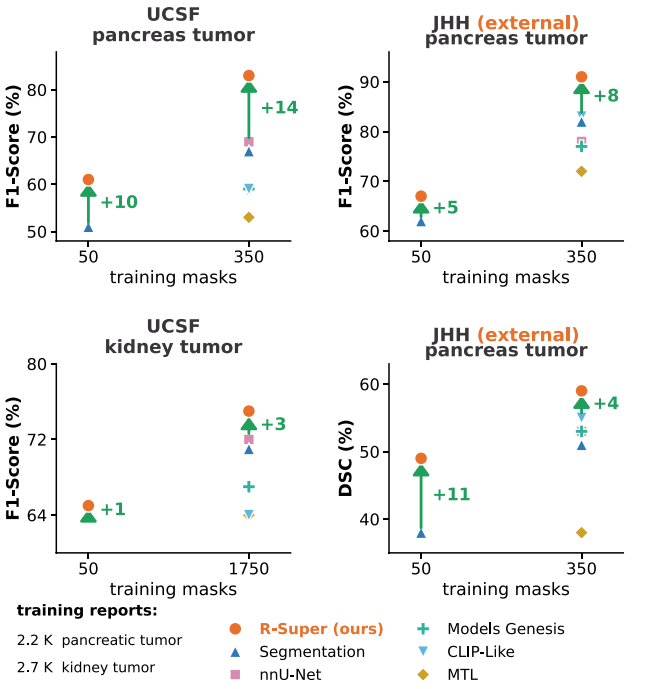

Fig. 3. By learning from reports, R-Super substantially outperforms six state-of-the-art methods on both internal (UCSF-Test) and external (JHH-Test) validation. The gain over standard segmentation (no reports) reaches up to +14% in tumor detection F1-Score and +11% DSC (green arrows). Each panel corresponds to one tumor type and one test set, and its x axis shows the number of training *masks* (50 from AbdomenAtlas-Subset; 344 or 1.7K from AbdomenAtlas 3.0). R-Super and the other report-based methods (CLIP, MTL, RG pseudo-labels) supplement these masks with 2.2K pancreatic and 2.7K kidney tumor *reports* from UCSF-Train. Both test sets are challenging: JHH-Test is out-of-distribution relative to training, and UCSF-Test spans a wide range of CT resolutions and contrast phases, including non-contrast scans. DSC is only reported for JHH-Test because UCSF-Test does not include ground-truth tumor masks, needed for calculating DSC.

the alignment between the ball center and the tumor center. The Gaussian decay and the weighted cross-entropy are formally defined in Appendix A.

4. Results

We present experiments at three scales of training data, using both public and private datasets. Section 4.1 trains on 6718 CT-Report pairs from a private dataset, combined with 50 to 1674 public CT-Mask pairs. Results are provided on both pancreatic and kidney tumors. This section also includes comprehensive comparisons against six alternative methods and ablation studies. Section 4.2 uses only public training data, demonstrating that R-Super can scale the largest public pancreatic tumor mask dataset (PanTS, 926 CT-Mask pairs) by adding CT-Report pairs from Merlin (1848 CT-Report pairs). Section 4.3 reaches the largest scale in this paper, with over 47,000 training CT scans. Section 4.4 evaluates the detection of pancreatic tumors on non-contrast CT without non-contrast tumor masks for training.

Training procedure. R-Super is a training strategy that can be applied to any segmentation architecture. In our experiments, we adopted MedFormer (Gao et al., 2022), a hybrid transformer-CNN architecture that has topped prior benchmarks (Bassi et al., 2024b). We kept its default training hyper-parameters (Appendix B). Because the competing approaches (Models Genesis, MTL, CLIP, report-guided pseudo-labels, and standard mask-based segmentation) are likewise training strategies rather than architectures, we paired each of them with the same MedFormer backbone and training hyper-parameters, so that performance differences reflect the training strategy rather than implementation choices. The only exception was nnU-Net (Isensee et al., 2021), which we ran with its self-configuring ResEncL architecture and automatic hyper-parameters. The MTL baseline additionally required a classification head, which we attached to the output of MedFormer's encoder.

Little training overhead and no inference overhead. Relative to a standard segmentation model using the same segmentation architecture (MedFormer), the R-Super training strategy increases training time by only 30% and does not change inference time. R-Super was trained for approximately 105 h on one NVIDIA H100 GPU, with a peak GPU memory usage of 46 GB. Memory usage can be reduced to a minimum of 23 GB per GPU by training on two GPUs with a batch size of one per GPU. Each report is processed by an LLM only once, before training; the LLM Llama 3 70B AWQ processes approximately 250 reports per hour, using vLLM on one NVIDIA RTX 6000 GPU. At inference, R-Super processes one CT scan in approximately 1.2 min on one NVIDIA H100 GPU, using only 10 GB of GPU memory. This modest memory requirement enables deployment on less powerful, widely available GPUs, supporting use in resource-constrained clinical settings. Due to its small training overhead and no test overhead, R-Super's training and inference times and memory usage are determined primarily by the underlying segmentation architecture, which can be arbitrarily chosen.

4.1. Reports improve performance with few or many masks

In our first experiment, we train R-Super with CT-Report pairs plus a varying number of masks, from 50 to 1674. We assess its performance

Table 1
R-Super outperforms state-of-the-art training methods on tumor detection and segmentation. On pancreatic tumors, R-Super achieves the highest DSC, NSD, F1-Score, and AUC across all baselines. On kidney tumors, it matches mask-based segmentation when few masks are available (50) and surpasses it when many are available (1.7 K). DSC and NSD are reported only in datasets where masks exist (UCSF-Test has none). Sensitivity (Se) and Specificity (Sp) correspond to the threshold that maximizes F1-Score. Statistical tests were performed to compare F1-Score and AUC (paired permutation test for F1-Score and DeLong's test for AUC). Orange highlights statistically significant gains ($p < 0.05$). Standard deviations and confidence intervals are available in the supplementary Table 6.

Train paradigm	Pancreas tumor												Kidney tumor					
	JHH-Test								UCSF-Test				UCSF-Test					
	mask	rep.	dsc	nsd	F1	AUC	Se	Sp	F1	AUC	Se	Sp	mask	rep.	F1	AUC	Se	Sp
few training masks (50)																		
standard segmentation	50	0	38	41	62	63	62	62	51	63	47	77	50	0	64	68	68	63
R-Super (ours)	50	2.2K	49	52	67	75	68	64	61	70	73	59	50	2.7K	65	66	73	55
medium / many training masks (344 / 1.7K)																		
CLIP-Like	344	2.2K	55	58	83	86	90	70	59	74	59	75	1.7K	2.7K	64	71	75	48
Multi-task learning	344	2.2K	38	43	72	80	78	60	53	62	65	50	1.7K	2.7K	64	71	68	63
Report-G Pseudo-labels	344	2.2K	56	62	80	79	80	78	69	83	65	86	1.7K	2.7K	71	74	80	60
Models Genesis	344	0	53	57	77	81	78	74	59	72	66	63	1.7K	0	67	73	77	54
nnU-Net	344	0	53	57	78	75	76	82	69	79	74	74	1.7K	0	72	74	87	53
standard segmentation	344	0	51	59	82	84	78	88	67	78	62	85	1.7K	0	71	73	65	68
R-Super (ours)	344	2.2K	**59**	**69**	**91**	**92**	94	88	**83**	**90**	83	89	1.7K	2.7K	**75**	**78**	80	70

on internal validation (on the same hospital that provided the CT-Report pairs) and external validation (a completely unseen hospital). The objective is to demonstrate that including many CT-Report pairs in the training set is useful whether few or many masks are available.

Datasets. (I) AbdomenAtlas 3.0 (Bassi et al., 2025b) is a large public CT dataset with 9262 CT–Mask pairs collected from 88 hospitals in 19 countries, including 344 pancreatic tumor CTs and 1674 kidney tumor CTs. **(II) AbdomenAtlas-Subset** is a subset of AbdomenAtlas 3.0 with 200 CT-Mask pairs, including 50 pancreatic tumor CTs, 50 kidney tumor CTs, and 100 normal CTs. **(III) UCSF-Train** is a large dataset (private) from UCSF and its affiliated institutions in California, USA. It has 6718 CT–Report pairs: 2229 pancreatic tumor CTs, 2738 kidney tumor CTs, and 1751 normal CTs. UCSF-Train does not include tumor masks. **(IV) UCSF-Test** is a test set drawn from the same hospital as UCSF-Train, including 169 kidney tumor CTs, 139 pancreatic tumor CTs, and 100 normal CTs. It was created by randomly splitting a dataset into training and testing (patient-level split). Therefore, UCSF-Test has a similar data distribution to UCSF-Train. UCSF-Test does not include tumor masks. **(V) JHH-Test** (Xia et al., 2022; Park et al., 2020) is an external test set from an unseen hospital, the Johns Hopkins Hospital (Maryland, USA), with 50 pancreatic tumor CTs and 50 normal CTs. No data from this hospital was seen by any of the AI models in training, making this dataset out-of-distribution. JHH-Test includes radiologist-drawn tumor masks. More details on the datasets are available in Fig. 2.

Training and Evaluation. We trained two configurations of R-Super. The first trained on UCSF-Train (many CT-Report pairs) plus AbdomenAtlas-Subset (few CT-Mask pairs). The second version trained on UCSF-Train plus AbdomenAtlas 3.0 (more CT-Mask pairs). We trained first on the CT-Mask pairs, followed by fine-tuning on CT-Mask plus CT-Report pairs. Evaluation was performed on UCSF-Test (internal) and JHH-Test (external).

R-Super improves tumor segmentation with few or many training masks. Fig. 3 compares R-Super against standard segmentation (trained without reports). Detailed results are provided in Table 1, examples of outputs in Fig. 4. Adding 2.2K pancreatic and 2.7K kidney tumor CT-Report pairs to training yielded consistent gains in F1-Score across every mask regime: up to +9.7% with few masks (50, AbdomenAtlas-Subset), up to +16% with a moderate number of masks (344, AbdomenAtlas 3.0), and +4.3% with many masks (1.7K, AbdomenAtlas 3.0). Similar gains are observed in DSC: +11.1% with 50 masks, and +4% with 350 masks. We only report DSC for pancreatic tumors, because only JHH-Test includes masks. In summary, by leveraging reports, R-Super can improve both tumor detection and segmentation, when few or many training masks are available.

R-Super generalizes to unseen hospitals. Fig. 3 shows that R-Super substantially outperforms standard segmentation on UCSF-Test (drawn from the same hospital as UCSF-Train) and on JHH-Test, a hospital never seen during training (out-of-distribution). Performance on JHH-Test was actually higher than on UCSF-Test, likely reflecting data differences: JHH-Test has high resolution (0.5 mm) and arterial-phase contrast, whereas UCSF-Test mixes a range of slice thicknesses and includes both contrast and non-contrast scans. On both test sets, R-Super surpasses six state-of-the-art methods, 3 of which also use reports during training (CLIP-Like, MTL, Report-Guided Pseudo-labels). The advantage is consistent across F1-Score, AUC, DSC, and NSD (Table 1).

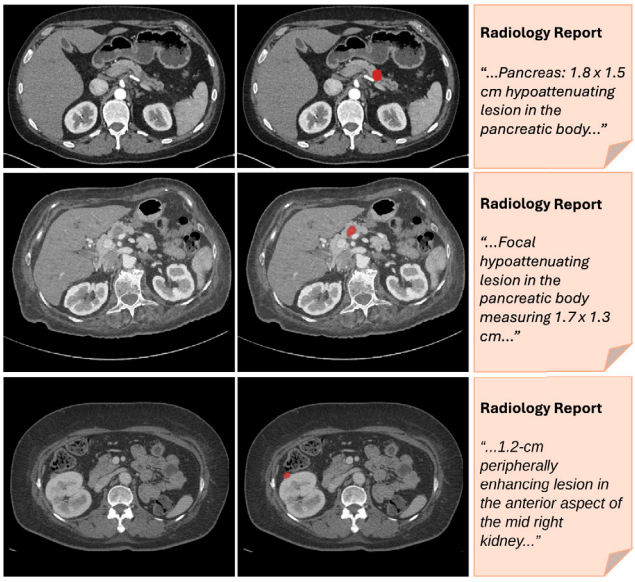

Fig. 4. Tumors segmented by R-Super. The examples show two cases of small pancreatic adenocarcinoma and one case of small renal carcinoma (bottom) correctly segmented by R-Super. Radiology reports are shown but they are not used in inference, only in training.

R-Super improves small tumor detection and segmentation. R-Super improved tumor detection and segmentation performance across the full size spectrum, including the small tumors (diameter $\leq$ 2 cm) that matter most for early cancer detection. Detailed results are presented in Supplementary Table 2. On small pancreatic tumors in the external JHH-Test, R-Super reached an F1-Score of 80 and AUC of 89, exceeding standard segmentation by +15 F1 and +6 AUC.

4.2. Reports scale the largest public pancreatic tumor segmentation dataset

In this experiment, we demonstrate that reports can scale the largest public pancreatic tumor dataset, PanTS (Li et al., 2025b). Reports were

Table 2
R-Super outperforms state-of-the-art methods on both small (diameter ≤ 2 cm) and large tumors. For pancreatic tumors (344 training masks), R-Super achieves the highest DSC, NSD, F1-Score, and AUC across all baselines, in both the small (≤ 2 cm) and large (> 2 cm) tumors. For kidney tumors (1.7 K training masks), R-Super leads in F1-Score and AUC on small tumors and in F1-Score on large tumors. DSC and NSD are reported only where masks exist (UCSF-Test has none). Sensitivity (Se) and Specificity (Sp) correspond to the threshold that maximizes F1-Score. Statistical tests were performed to compare F1-Score and AUC (paired permutation test for F1-Score and DeLong's test for AUC). Orange highlights statistically significant gains ($p < 0.05$). Test set sizes: JHH-Test contains 15 small and 35 large pancreatic tumors; UCSF-Test contains 90 small and 28 large pancreatic tumors (21 CTs with tumors of unknown size were excluded from this table) and 77 small and 65 large kidney tumors (27 excluded). Tumor sizes were extracted from reports in UCSF-Test, and from masks in JHH-Test. Standard deviations and confidence intervals are available in the supplementary Table 7.

Train paradigm	Pancreas tumor												Kidney tumor					
	JHH-Test								UCSF-Test				UCSF-Test					
	mask	rep.	dsc	nsd	F1	AUC	Se	Sp	F1	AUC	Se	Sp	mask	rep.	F1	AUC	Se	Sp
small tumors (diameter ≤ 2 cm)																		
few training masks (50)																		
standard segmentation	50	0	7	17	38	61	53	62	41	59	40	77	50	0	43	59	53	63
R-Super (ours)	50	2.2K	15	25	35	68	47	64	54	72	76	59	50	2.7K	47	61	65	55
medium / many training masks (344 / 1.7K)																		
CLIP-Like	344	2.2K	11	19	68	**90**	100	71	50	74	54	75	1.7K	2.7K	40	65	57	48
Multi-task learning	344	2.2K	15	26	54	83	87	60	42	61	60	50	1.7K	2.7K	46	65	57	63
Report-G Pseudo-labels	344	2.2K	19	32	61	77	73	80	63	82	62	86	1.7K	2.7K	50	71	66	60
Models Genesis	344	0	10	20	62	85	80	76	48	70	60	63	1.7K	0	43	65	60	54
nnU-Net	344	0	7	17	55	74	60	82	60	78	70	74	1.7K	0	53	71	79	53
standard segmentation	344	0	17	36	65	83	67	88	59	77	58	85	1.7K	0	39	69	44	68
R-Super (ours)	344	2.2K	**25**	**48**	**80**	89	93	88	**75**	**90**	77	89	1.7K	2.7K	**56**	**78**	69	69
large tumors (diameter > 2 cm)																		
few training masks (50)																		
standard segmentation	50	0	3	5	59	63	65	62	35	78	61	77	50	0	56	75	82	63
R-Super (ours)	50	2.2K	8	13	67	78	76	64	31	75	79	59	50	2.7K	54	72	86	55
medium / many training masks (344 / 1.7K)																		
CLIP-Like	344	2.2K	13	19	75	84	85	71	43	84	82	75	1.7K	2.7K	54	77	95	48
Multi-task learning	344	2.2K	18	28	63	78	74	60	28	70	82	50	1.7K	2.7K	54	77	78	63
Report-G Pseudo-labels	344	2.2K	19	31	78	79	82	80	55	90	82	86	1.7K	2.7K	62	75	97	60
Models Genesis	344	0	13	21	72	79	76	76	36	83	86	63	1.7K	0	58	**80**	97	54
nnU-Net	344	0	11	19	79	75	82	82	46	89	93	74	1.7K	0	57	76	97	53
standard segmentation	344	0	18	33	82	84	82	88	56	88	86	85	1.7K	0	63	75	91	68
R-Super (ours)	344	2.2K	**33**	**51**	**89**	**93**	94	88	**68**	**94**	96	89	1.7K	2.7K	**67**	75	97	69

Table 3
R-Super sets a new state-of-the-art on the largest public pancreatic tumor dataset. Combining CT-Mask pairs from PanTS (the largest public pancreatic tumor CT-Mask dataset) with CT-Report pairs from Merlin (the largest public abdominal CT-Report dataset), R-Super outperforms a segmentation model trained on CT-Mask pairs alone (PanTS) by +7% sensitivity on small pancreatic tumors. R-Super also surpasses all alternative models on AUC, F1-Score, DSC, and NSD. All models were trained on the same 926 CT-Mask pairs from PanTS. Multi-task learning and R-Super additionally trained on 1848 CT-Report pairs from Merlin. DSC and NSD are reported only where masks exist; Merlin-Test has none. Sensitivity (Se), Specificity (Sp), and F1-Score correspond to the threshold that maximizes F1-Score for each model. Statistical tests were performed to compare R-Super and standard segmentation for F1-Score and AUC (paired permutation test for F1-Score and DeLong's test for AUC); orange highlights statistically significant gains ($p < 0.05$). Standard deviations and confidence intervals are available in the supplementary Table 8.

Train paradigm	Training		JHH-Small						JHH-Large						Merlin-Test			
	mask	rep.	dsc	nsd	F1	AUC	Se	Sp	dsc	nsd	F1	AUC	Se	Sp	F1	AUC	Se	Sp
nnU-Net	926	0	33	58	83	91	80	87	49	72	88	93	84	92	69	76	68	73
Multi-task learning	926	1,848	33	57	92	95	92	92	50	73	97	99	98	96	69	75	63	82
standard segmentation	926	0	31	55	91	94	86	98	50	73	**99**	**100**	98	100	67	74	60	82
R-Super (ours)	926	1,848	**34**	**59**	**95**	**97**	93	98	**52**	**74**	98	**100**	100	96	**78**	**83**	77	79

sourced from the public Merlin dataset (Blankemeier et al., 2024). Therefore, this experiment trains on only public data, ensuring reproducibility. Here, we compare R-Super against standard mask-only segmentation, the nnU-Net, and multi-task learning.

Datasets. (I) PanTS (Li et al., 2025b) is the largest public CT-Mask dataset for pancreatic tumors, with 9901 CT-Mask pairs from 143 medical institutions in 17 countries, including 1077 with pancreatic tumors. For our experiments, we augmented PanTS with **(II) Merlin-Train**. Merlin Blankemeier et al. (2024) is the largest public abdominal CT dataset, with 25,494 CT-Report pairs (no masks). In Merlin-Train, we included 1848 pancreatic tumor CT-Report pairs from Merlin, plus 1848 randomly selected no tumor CT-Report pairs. **(III) Merlin-Test** includes the remaining 200 pancreatic tumor CT-Report pairs in Merlin, plus 200 no tumor CT-Report pairs. Merlin-Test is an internal test set. For external validation, we test on **(IV) JHH-Large**, a set of 50 large (> 2 cm) pancreatic tumor CT-Mask pairs and 50 no tumor CT-Mask pairs. Additionally, we test on **(V) JHH-Small**, which includes 86 small pancreatic tumor CT-Mask pairs plus 86 no tumor CT-Mask pairs. More details on the datasets are available in Fig. 2.

Training and Evaluation. We trained R-Super first on the CT-Mask pairs from PanTS, followed by fine-tuning on CT-Mask (PanTS) plus CT-Report pairs (Merlin-Train). Evaluation was performed on Merlin-Test (internal), JHH-Large (external), and JHH-Small (external). JHH-Small is highly challenging, as small tumors are difficult to detect.

R-Super scales the largest public pancreatic tumor segmentation dataset. Results for the 3 test sets are shown in Table 3. By leveraging reports, R-Super achieved +11% higher F1-Score and +9% higher AUC over training on masks only (standard segmentation), testing on Merlin-Test (internal). Even on external testing on small tumors (JHH-Small), R-Super surpassed standard segmentation by +4% F1-Score, +3% AUC, and +3% DSC. As in Section 4.1, R-Super also surpassed the alternative training methods (Table 3). The results in this section demonstrate that even the largest collections of CT-Mask pairs can be effectively scaled with CT-Report pairs, improving inter-hospital generalization and improving the detection and segmentation of early tumors.

Table 4
R-Super improves tumor detection in non-contrast CT and surpasses mask-based training when large amounts of both reports and masks are available. On Swiss-Test, an external dataset which we are publicly releasing, R-Super outperforms standard mask-based segmentation by +8% sensitivity at the same specificity, for pancreatic tumor detection. On UCSF-Test-Non-Contrast, an internal test set of non-contrast CT scans, R-Super outperformed the standard segmentation model by +19% sensitivity at the same specificity. DSC and NSD are absent because there are no ground-truth tumor masks for testing. All models were trained on the same 3488 pancreatic tumor CT-Mask pairs from JHH (contrast-enhanced); Multi-task learning and R-Super additionally trained on 28,295 tumor CT-Report pairs from UCSF-Huge (41,418 CT-Report pairs, 4809 non-contrast). Sensitivity (Se), Specificity (Sp), and F1-Score correspond to the threshold that maximizes F1-Score for each model. Statistical tests compare R-Super and standard segmentation (paired permutation test for F1-Score, DeLong's test for AUC); orange highlights statistically significant gains ($p < 0.05$). Standard deviations and confidence intervals are available in the supplementary Table 9.

Train paradigm	Training		Swiss			
	mask	rep.	F1	AUC	Se	Sp
Swiss-Test						
nnU-Net	3,488	0	78	74	74	71
Multi-task learning	3,488	28,295	74	76	64	85
standard segmentation	3,488	0	80	82	72	86
R-Super (ours)	3,488	28,295	**85**	**85**	80	86
UCSF-Test-Non-Contrast						
nnU-Net	3,488	0	52	68	67	73
Multi-task learning	3,488	28,295	19	54	10	99
standard segmentation	3,488	0	62	74	50	96
R-Super (ours)	3,488	28,295	**75**	**92**	69	96

4.3. Reports scale the largest private pancreatic tumor segmentation dataset

In this experiment, we explore the largest scale in this paper, both in terms of masks and reports: 6212 CT-Mask pairs and 41,418 CT-Report pairs. To reach this scale, we use private training data. However, we make our test set public, contributing a new CT-Report dataset (Swiss-Test). Again, we compare R-Super against standard mask-only segmentation, nnU-Net, and multi-task learning.

Datasets. JHH (Xia et al., 2022) is, to the best of our knowledge, the largest private CT-Mask dataset for pancreatic tumor segmentation. It includes 6212 CT-Mask pairs, of which 3488 have pancreatic tumors, of which 1093 are small (≤ 2 cm). **UCSF-Huge** includes 41,418 CT-Report pairs, of which 28,295 have pancreatic tumors, of which 11,413 are small. **Swiss-Test** is a new dataset, sourced from the University Hospital Basel (Basel, Switzerland). This dataset includes 684 CT-Report pairs, of which 445 include pancreatic tumors. Here, we leverage it as an external test set. We will publicly release this dataset. More details on the datasets are available in Fig. 2.

Training and Evaluation. We trained R-Super first on the CT-Mask pairs from JHH, followed by fine-tuning on CT-Mask (JHH) plus CT-Report pairs (UCSF-Huge). Evaluation was performed on Swiss-Test.

R-Super scales the largest private pancreatic tumor segmentation dataset. Table 4 shows the performance of R-Super on the Swiss-Test external dataset. R-Super surpassed standard mask-based segmentation, achieving +8% higher pancreatic tumor detection sensitivity at matched specificity. This result demonstrates that reports can improve the performance of tumor detection and segmentation even when large CT-Mask datasets exist. Notably, the CT-Report pairs in UCSF-Huge include all pancreatic tumor cases registered in the UCSF hospital and affiliated Californian institutions in the last 28 years. Downloading and preparing all these CT-Report pairs took around one month and incurred no monetary cost. In contrast, creating the 3488 pancreatic tumor masks in JHH took 5 years, with the support of 8 radiologists, and millions in funding. This demonstrates that reports are an accessible path to scale tumor segmentation datasets and improve AI performance.

4.4. Reports enable pancreatic tumor detection on non-contrast ct

In this experiment, we evaluate R-Super for pancreatic tumor detection on non-contrast CT scans, without training on non-contrast tumor masks. We use the same model evaluated in the previous section: R-Super trained on JHH (6212 CT-Mask pairs) plus UCSF-Huge (41,418 CT-Report pairs). JHH includes only contrast-enhanced CT-Mask pairs. UCSF-Huge includes contrast-enhanced CT-Report pairs and 4809 non-contrast CT-Report pairs, 3106 with pancreatic tumors. We compare R-Super against the same baselines as in the previous section: standard mask-only segmentation, nnU-Net, and multi-task learning. We test on **UCSF-Test-Non-Contrast**, a dataset with 207 non-contrast CT-Report pairs from UCSF, including 48 CT-Report pairs with pancreatic tumors, 15 with only small tumors (≤ 2 cm).

R-Super detects pancreatic tumors on non-contrast CT scans. R-Super surpassed the standard segmentation model by +19% in sensitivity at matched specificity for detecting malignant pancreatic tumors on non-contrast CT scans (Table 4). We trained on 3106 non-contrast CT-Report pairs with pancreatic tumors. Performance may improve even further with more non-contrast CT-Report pairs for training. Seeing pancreatic tumors in CT scans taken without intravenous contrast is considered extremely challenging for radiologists (Cao et al., 2023). Consequently, drawing pancreatic tumor masks on non-contrast CT scans is also extremely difficult, making standard mask-supervised segmentation challenging. Previous closed-source work addressed this challenge by asking radiologists to draw a large number of masks on contrast-enhanced CT scans and registering these masks to paired non-contrast CT scans from the same patients (Cao et al., 2023). Conversely, we did not require drawing any mask for the patients with non-contrast CT. To train R-Super, we only needed to supplement an existing dataset of contrast-enhanced CT-Mask pairs (JHH) with a large-scale CT-Report dataset (UCSF-Huge), which includes non-contrast and contrast-enhanced CT-Report pairs. Report supervision therefore allowed us to extend pancreatic tumor segmentation models and datasets from contrast-enhanced to non-contrast CT scans without additional masks.

4.5. Ablation studies

Table 5 shows ablation studies performed with R-Super. Their aim was to understand the importance of each R-Super loss function, the resistance of R-Super to LLM errors or report errors in training, the importance of the LLM size, the importance of organ masks, and whether training R-Super to segment a single tumor type surpasses training a generalist R-Super to segment multiple tumor types. Ablation studies were conducted with the same training and testing configuration presented in Section 4.1, with AbdomenAtlas 3.0 and UCSF-Train training sets, and UCSF-Test and JHH-Test test sets.

Training with the Volume Loss alone outperforms the Ball Loss alone, and training with both losses together provides the best performance. The Ball Loss is more precise than the Volume Loss, by optimizing segmented tumors to match more details from the radiology report. However, the Ball Loss is harder to optimize when used alone. This explains why training with the Volume Loss alone surpassed training with the Ball Loss alone (Table 5). The Ball Loss identifies the most likely tumor locations based on both the radiology report and the segmentation model's current output. It then encourages the model's output to better match these tumor locations. Early in training, the model's outputs are inaccurate, making the estimated tumor locations noisy and slowing training convergence. Importantly, this noise does not systematically direct the model toward incorrect segmentations. The Ball Loss may temporarily reinforce a false positive in an individual CT scan, but the same false positive is not reinforced consistently across

Table 5
Ablation studies of R-Super. *Loss functions:* Combining the Volume Loss and Ball Loss achieved the best performance. Training with the Volume Loss alone surpassed training with the Ball Loss alone, because the Volume Loss is easier to optimize. *LLM errors (or report errors):* R-Super remained highly accurate when errors were injected into up to 10% of the tumor reports in the training dataset. These errors can simulate either errors in the report or errors in the LLM analysis of the report. *LLM size:* training with small LLMs (Qwen 3 4B or 0.6B instead of Llama 3 70B) substantially reduced R-Super performance. *Organ mask quality:* organ masks generated by an nnU-Net trained on only 50 CT scans were sufficient to accurately train R-Super. Even when organ masks were fully removed, R-Super still outperformed the standard segmentation model (Table 1). *Report information removed:* tumor diameters are important for R-Super, particularly for DSC. Standard deviations and confidence intervals are available in the supplementary Table 10. Results are reported at the operating point of maximum F1-Score for all models.

train paradigm	Pancreas tumor												Kidney tumor					
	JHH-Test								UCSF-Test				UCSF-Test					
	mask	rep.	dsc	nsd	F1	AUC	Se	Sp	F1	AUC	Se	Sp	mask	rep.	F1	AUC	Se	Sp
R-Super (ours)	344	2.2 K	59	69	91	92	94	88	83	90	83	89	1.7 K	2.7 K	75	78	80	70
loss functions																		
Volume Loss only	344	2.2 K	59	61	94	97	90	98	76	89	80	82	1.7 K	2.7 K	75	74	86	63
Ball Loss only	344	2.2 K	59	59	83	92	78	90	71	82	64	90	1.7 K	2.7 K	74	75	87	58
LLM size																		
small LLM (4B)	344	2.2 K	63	74	82	84	76	92	70	80	74	79	1.7 K	2.7 K	72	76	89	47
tiny LLM (0.6B)	344	2.2 K	41	50	74	80	71	80	63	73	54	90	1.7 K	2.7 K	73	76	86	55
report information removed																		
no pancreas subsegment	344	2.2 K	58	69	85	90	80	92	71	81	73	81	1.7 K	2.7 K	72	75	86	54
no tumor size	344	2.2 K	50	59	85	87	76	98	67	76	64	84	1.7 K	2.7 K	73	76	88	55
organ mask importance																		
50 CT organ segmenter	344	2.2 K	62	74	95	97	96	94	76	85	86	77	1.7 K	2.7 K	72	76	84	58
no organ mask	344	2.2 K	60	69	88	93	80	98	69	81	68	83	1.7 K	2.7 K	73	76	85	58
injected LLM errors (or report errors)																		
+5% errors	344	2.2 K	62	74	93	94	90	96	72	85	73	84	1.7 K	2.7 K	74	77	91	51
+10% errors	344	2.2 K	59	71	90	95	88	92	78	87	78	87	1.7 K	2.7 K	73	77	82	63
+20% errors	344	2.2 K	61	73	89	90	86	94	71	82	76	79	1.7 K	2.7 K	72	76	86	53
100% errors	344	2.2 K	53	61	70	74	59	90	56	68	72	51	1.7 K	2.7 K	66	70	77	50

patients. E.g., similar false positives are penalized when appearing in no-tumor scans. In contrast, learning to segment real tumors (true positives) allows the segmentation model to consistently match the tumor locations, counts, and sizes described in the reports, consistently minimizing the Ball Loss. Thus, training with the Ball Loss and the Volume Loss together surpasses training with the Volume Loss alone. The less noisy Volume Loss stabilizes early training, while the Ball Loss refines the segmentation model's outputs and becomes increasingly precise as the model's predictions improve.

On the relative contribution of the Ball Loss. We analyze where training with the Ball Loss plus the Volume Loss (standard R-Super) has the greatest advantage over training with the Volume Loss only. To this end, we compare R-Super to the Volume Loss only ablation on several subsets of UCSF-Test: small tumors, multiple tumors in one CT (multifocal disease), and complex tumors. Since UCSF-Test is larger than JHH-Test, it affords more statistical power for these subset analyses. Adding the Ball Loss to the Volume Loss significantly improves overall tumor detection for both organs: pancreatic tumors gain +3 AUC points (95% CI $[+1, +6]$, n=139) and kidney tumors +2 ($[+1, +4]$, n=169). Two subsets gain well above this average. **Multifocal disease**: scans with $\geq$ 2 kidney tumors gain +3 AUC points ($[+1, +4]$, n=85), above the +2 kidney average. For pancreas, multifocal lesions are scarce (18), making this subset analysis not significant. **Anatomically complex pancreatic tumors**: tumors in the pancreatic tail and iso-attenuating tumors are rare and hard to detect, and gain +6 AUC points ($[+1, +12]$, n=27), double the +3 pancreatic average. **Small tumors**: tumors ≤ 2 cm gain in line with the overall average, +3 AUC points for pancreas ($[-1, +6]$, n=90) and +2 for kidney ($[-1, +5]$, n=77). In short, the Ball Loss is not redundant with the Volume Loss; its improvements are largest precisely on the multifocal and complex tumors that are hard to detect.

R-Super tolerates LLM or report errors. R-Super uses a large LLM (Llama 3.1 70B AWQ) to extract tumor information from reports accurately. A radiologist evaluated this LLM on 447 reports, and the LLM reached 96% accuracy (Bassi et al., 2025b). To quantify how resistant R-Super is to LLM errors or report errors, we performed ablation studies injecting increasing numbers of synthetic errors into the information that the LLM extracted from reports. Table 5 shows that adding synthetic errors to 5% or 10% of the tumor CT-Report pairs in the training dataset did not substantially change the R-Super performance. E.g., it only reduced tumor detection AUC by at most -5%. By adding +20% of synthetic errors, AUC dropped at most -8%. Still, the performance of R-Super remained above the standard segmentation model (Table 1). With 100% of synthetic errors, R-Super performance dropped massively (at most -22% AUC), performing below the standard segmentation model (trained on masks only). This is expected, as a 100% error rate turns report supervision essentially into training noise. The synthetic errors simulated, in equal proportions, four types of mistakes that LLMs could make when extracting tumor information from reports: tumor location errors were simulated by switching the pancreatic sub-segment for pancreatic tumors and switching the kidney laterality for kidney tumors; tumor size errors were simulated by replacing a tumor's size with the size of another tumor in the dataset; false positives were simulated by copying extracted tumor information from one report and adding it as an extra tumor to another report; and false negatives were simulated by deleting the information of an LLM-extracted tumor.

Large LLMs outperform small LLMs. Besides injecting synthetic errors, we also trained R-Super with smaller LLMs, which should produce more real-world LLM errors. R-Super originally used a large LLM (Llama 3 70B AWQ). In our ablations, we replaced it with Qwen 3 4B FP8 or Qwen 3 0.6B FP8 (Yang et al., 2025). Training with these smaller LLMs substantially reduced R-Super performance. Training with Qwen 3 4B, R-Super performed similarly to training with Llama 3 70B plus 20% synthetic errors. In most metrics, it slightly outperformed the standard segmentation model. With Qwen 3 0.6B, R-Super was often outperformed by the standard segmentation model (Table 5). These results suggest that current small LLMs may make too many errors in extracting tumor information from reports, making report supervision unreliable. We therefore strongly recommend large LLMs. Notably, the LLM only needs to run once per report, not incurring an excessive computational cost (Llama 3 70B AWQ can analyze 5000 reports in less than 1 day with a single NVIDIA RTX6000 GPU).

R-Super can work without organ masks. The R-Super loss functions are guided by organ masks. We dilate these masks to compensate for organ segmentation errors and for tumors growing beyond organ borders. Originally, the organ masks were generated by an nnU-Net trained on the full AbdomenAtlas 3.0 dataset of 9262 CT scans. In our ablation, we replaced this nnU-Net with an nnU-Net trained on only 50 CT scans, randomly selected from AbdomenAtlas 3.0. This replacement did not substantially reduce R-Super's performance. E.g., tumor detection AUC decreased by at most -5% (Table 5). To evaluate the quality of these nnU-Nets in organ segmentation, we tested them on the JHH-Test dataset, which includes manually annotated pancreas and kidney masks (but not pancreas sub-segment masks). The nnU-Net trained on the full AbdomenAtlas 3.0 achieved a mean DSC of 81.8% (SD: 12.6%) for the pancreas and 98.0% (SD: 1.0%) for the kidneys. In comparison, the nnU-Net trained on only 50 CT scans achieved a mean DSC of 81.6% (SD: 10.9%) for the pancreas and 97.4% (SD: 2.5%) for the kidneys. In a more radical ablation, we fully removed the organ mask guidance from the R-Super loss functions. Without organ masks, the Volume Loss estimated the segmented tumor volume from the entire segmentation model output, and the Ball Loss searched the entire segmentation model output for the most likely tumor locations. Without organ masks, tumor detection AUC decreased by at most -9% (Table 5), but R-Super still consistently outperformed a standard tumor segmentation model (Table 1). In conclusion, organ masks make the R-Super loss functions more precise and improve performance. However, these organ masks can be created by organ segmentation models trained on few images, and even without organ masks, the R-Super loss functions can still improve performance.

Tumor diameters are a strong training signal. To assess the contribution of different information in radiology reports to the R-Super performance, we independently removed tumor diameters or tumor sub-segments from all training reports. Removing tumor sub-segments decreased tumor detection AUC by up to −9% and DSC by −1%. In contrast, removing tumor diameters decreased AUC by up to −14% and DSC by −9% (Table 5). Thus, tumor diameters and sub-segments are useful, but tumor diameters provide stronger supervision to segmentation.

Training R-Super to segment a single tumor type performs slightly better than training it to segment two tumor types. Training one R-Super model for pancreatic tumors only (rather than a single model for both pancreatic and kidney tumors) raised pancreatic tumor performance to DSC/NSD/Se/Sp/F1/AUC of 62/64/88/96/92/98 on JHH-Test and 86/86/83/91 on UCSF-Test, but it requires training one model per tumor type, which scales poorly as more tumor types are added.

5. Conclusion

R-Super transforms radiology reports into supervision for tumor segmentation, allowing AI performance to scale with the vast number of CT-Report pairs already stored in hospitals and increasingly available in public datasets (Hamamci et al., 2024; Bassi et al., 2025b; Blankemeier et al., 2024; Li et al., 2026a,b). This work demonstrates that reports substantially improve both tumor detection and segmentation across multiple data scales, from adding 6000 CT-Report pairs to a small set of 100 CT-Mask pairs, up to adding over 40,000 CT-Report pairs (decades of pancreatic cancer cases) to over 6000 CT-Mask pairs (a dataset that took 8 radiologists 5 years to create).

We also demonstrated that adding non-contrast CT-Report pairs to a contrast-enhanced CT-Mask dataset enables pancreatic tumor detection on non-contrast CT, without non-contrast tumor masks for training. This is valuable because tumors are often extremely difficult to see on non-contrast CT, even for expert radiologists. Therefore, non-contrast tumor masks are exceptionally scarce and currently unavailable in public pancreatic tumor datasets. R-Super helps overcome the need for non-contrast tumor masks.

For reproducibility, we include not only experiments with private data, but also with public data. Furthermore, we release a new public CT-Report test set (Swiss-Test). Notably, reports enable easier merging of data across hospitals, without the need to create masks for each hospital. This improves training data diversity, improving generalization to unseen hospitals, as shown by consistent performance gains on multiple external test sets (JHH-Test, Merlin-Test, Swiss-Test).

R-Super also surpassed alternative methods of learning from reports. For instance, R-Super outperformed CLIP-style pretraining and multi-task learning, trained on the same datasets. This is possible because R-Super uses reports to supervise tumor segmentation directly, rather than using reports to supervise an auxiliary task. To minimize the R-Super losses, the segmentation model learns to segment tumors that match tumor descriptions in reports.

The scales reached in this study are already substantial. For instance, our largest training set (Section 4.3) contains over 30 times more pancreatic tumor cases than the largest public pancreatic tumor segmentation dataset. Yet new public datasets and multi-institution collaborations can extend this scale much further. In other fields, such as natural language processing and computer vision, major performance gains came from learning at scale from readily available text. This study suggests that the same path is possible for tumor detection and segmentation.

CRediT authorship contribution statement

Pedro R.A.S. Bassi: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Wenxuan Li:** Writing – review & editing, Visualization, Validation, Investigation, Formal analysis, Data curation. **Jakob Wasserthal:** Validation, Resources, Data curation. **Jieneng Chen:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Xinze Zhou:** Validation, Investigation. **Zheren Zhu:** Data curation. **Chuntung Zhuang:** Validation, Investigation. **Sergio Decherchi:** Writing – review & editing, Supervision, Funding acquisition. **Andrea Cavalli:** Writing – review & editing, Supervision. **Kang Wang:** Writing – review & editing, Visualization, Supervision, Funding acquisition, Data curation. **Yang Yang:** Writing – review & editing, Supervision, Funding acquisition, Data curation. **Alan Yuille:** Writing – review & editing, Supervision, Resources, Investigation, Funding acquisition. **Zongwei Zhou:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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.

Appendix A. Details on the loss functions

A.1. Weighted cross-entropy and Gaussian ball kernel

Weighted cross-entropy. We replace the cross-entropy in Eq. (14) with a rank-weighted variant that down-weights voxels the model is less confident about. Tumor borders typically receive intermediate predicted

probabilities, so this rank-weighting reduces the gradient contribution of border voxels and concentrates supervision on confidently predicted tumor voxels.

This cross-entropy compares the tumor segmentation output $\mathbf{T}^o$ with the binary mask $\mathbf{M}$, which marks the most probable tumor voxels in organ o (Eq. (17)). We sort the voxels in $\mathbf{M}$ by predicted tumor probability in descending order, assigning each voxel a rank $r(x,y,z) \in \{0,\dots,n-1\}$, where $r=0$ is the highest-probability voxel and n is the total number of voxels in $\mathbf{M}$. The weight w then decays geometrically with rank:

$$\begin{aligned}&\mathrm{CE}\big((1-\mathbf{M}')\odot\mathbf{T}^o,\,\mathbf{M}\big)=\\&-\tfrac{1}{|\Omega|}\textstyle\sum_{(x,y,z)\in\Omega} w(x,y,z)\,\Big[\,\mathbf{M}\log\big((1-\mathbf{M}')\mathbf{T}^o\big)\\&+(1-\mathbf{M})\log\big(1-(1-\mathbf{M}')\mathbf{T}^o\big)\,\Big]\end{aligned}\tag{18}$$

$$w(x,y,z)=\begin{cases}\dfrac{n\,(1-c)^{r(x,y,z)/n}}{\sum_{r'=0}^{n-1}(1-c)^{r'/n}}, & \text{if } \mathbf{M}(x,y,z)=1\\ 1, & \text{otherwise}\end{cases}\tag{19}$$

The rank-weighting applies only to voxels that are 1 in $\mathbf{M}$; all background voxels (0) keep weight 1, so the loss outside the tumor region in $\mathbf{M}$ is unchanged from standard cross-entropy. The multiplication of $\mathbf{T}^o$ by $(1-\mathbf{M}')$ zeroes the gradient at the margin voxels, as in Eq. (14). Ω is the set of all voxels in $\mathbf{T}^o$. The hyperparameter $c \in (0,1)$ controls how sharply the weight concentrates on the top-ranked voxels: we set $c=0.5$, so the most confident half of the voxels in $\mathbf{M}$ carry half of the total tumor-voxel weight. The normalization in Eq. (19) ensures $\sum_{(x,y,z)\in\mathbf{M}} w(x,y,z) = n$, matching the scale of the unweighted cross-entropy; setting $c \to 0$ recovers the standard cross-entropy. This weighting scheme is based on Global Weighted Rank Pooling (Kolesnikov and Lampert, 2016), adding a hard cutoff at n.

Gaussian-decay ball kernel. In the Ball Loss, Ball Convolutions locate each tumor i (Eq. (9)). The kernel of the Ball Convolution is a ball with the same diameter as tumor i, according to the report (Eq. (7)). The Ball Convolution is applied at the output of the segmentation model (per-voxel tumor probabilities), excluding tumor probabilities outside organ o ($\mathbf{T}^o \odot \mathbf{O}$, where $\mathbf{O}$ is the organ segmentation mask). At each position of this kernel, it sums all the per-voxel tumor probabilities inside the ball. With a uniform binary kernel, every voxel inside the ball contributes equally to this sum, so the convolution maximum is insensitive to where, within the ball, predicted tumor probability is highest. We want the argmax to favor positions where probability is concentrated near the ball center, since predicted tumor probabilities tend to peak at tumor centers.

We achieve this by replacing the binary kernel $\mathbf{k}_i$ with a Gaussian-tapered version: largest at the kernel center, smaller near the ball boundary, zero outside the ball:

$$\mathbf{k}_i(u,v,w)=\mathbb{1}\Big[\sqrt{u^2+v^2+w^2}\le\tfrac{d_i}{2}\Big]\,e^{-\frac{u^2+v^2+w^2}{2\sigma_i^2}}\tag{20}$$

$$\text{where } \sigma_i=\alpha\,\frac{d_i}{2},\qquad \alpha=3\tag{21}$$

Here (u,v,w) are the kernel coordinates measured from the kernel center, and d_i is the diameter reported for tumor i. The indicator $\mathbb{1}[\cdot]$ preserves the hard cutoff at the ball boundary from Eq. (7), so $\mathbf{k}_i$ is zero outside the ball. The Gaussian standard deviation σ_i is scaled to the ball radius $d_i/2$ via α, so the decay shape is identical across reported diameters. The normalization is chosen so $\sum_{u,v,w}\mathbf{k}_i(u,v,w)=1$, keeping Ball Convolution outputs comparable across kernel sizes.

We set $\alpha=3$, which gives a boundary-to-center kernel ratio of $\exp(-1/18)\approx 0.95$. The decay is therefore very mild: enough to break argmax ties in favor of positions where predicted probability is concentrated near the kernel center, but small enough that the output of the Ball Convolution is close to a sum of the tumor probabilities inside each ball position. The Gaussian decay is designed to improve the Ball Loss centering, but it is not necessary for the loss to perform well.

A.2. Volume loss for reports without tumor size or count

The standard Volume Loss penalizes the segmented tumor volume $V_{s,o}$ for deviating from the reported volume $V_{r,o}$, which is computed from per-tumor diameters. When a report states that an organ contains tumors but does not provide diameters or counts, $V_{r,o}$ cannot be computed, so the standard Volume Loss cannot be applied. In this case, we replace the precise $V_{r,o}$ target with a tolerated range $[V_{min}, V_{max}]$. Any segmented volume inside this range incurs zero loss and zero gradient, while volumes outside the range are penalized as before.

We calibrate V_{min} and V_{max} from the diameter distribution in our reports: fewer than 5.5% of images have all tumors below 5 mm and fewer than 1.66% have tumors above 120 mm. We therefore set $V_{min} = 65\ \mathrm{mm}^3$ and $V_{max} = 904,779\ \mathrm{mm}^3$, the volumes of spheres at these two diameters. The high-tolerance loss (Eq. (22)) implements the tolerated range by replacing $V_{r,o}$ in the standard L_{forg} (Eq. (4)) with a target $\widehat{V}_{r,o}$ that equals $V_{s,o}$ whenever $V_{s,o}$ lies in $[V_{min}, V_{max}]$. With $\widehat{V}_{r,o} = V_{s,o}$, the loss and its gradient vanish; outside the range, $\widehat{V}_{r,o}$ pulls the segmented tumors' volume back toward the range.

$$L_{\text{vol},o}=L_{\text{forg},o}\Big(V_{s,o},\,\widehat{V}_{r,o}\Big)+L_{\text{bkg},o}(\mathbf{T}^o)\tag{22}$$

$$\text{where}\quad \widehat{V}_{r,o}=\begin{cases}V_{min} & \text{if } V_{s,o}<V_{min},\\ V_{s,o} & \text{if } V_{min}\le V_{s,o}\le V_{max},\\ V_{max} & \text{if } V_{s,o}>V_{max}.\end{cases}\tag{23}$$

A.3. Ball loss for reports without tumor size or count

The standard Ball Loss requires reports to provide per-tumor diameters because diameters are used to define the kernel of the Ball Convolution, and because the loss derives the voxel count N_i from the reported diameters (Section 3.3). The high-tolerance variant of the Ball Loss accepts reports with missing diameters or missing tumor counts by enforcing only a lower bound on the segmentation: the model must segment at least one small tumor in organ o, but it is not penalized for segmenting a larger tumor or multiple tumors.

First, we locate the most probable voxels for one small tumor by running the Ball Convolution with a small diameter (we use 5 mm, the same lower cutoff used in the high-tolerance Volume Loss above). This produces a small highest-probability ball, which may actually be inside of a large tumor. Inside this ball, we select the N_i most probable voxels (with N_i set to the voxel count of a 5 mm sphere, reduced by 20%). We then create a mask $\mathbf{M}$ that is 1 in these N_i most probable voxels and 0 outside, as in the standard Ball Loss (Eq. (12)).

The mask $\mathbf{M}'$, originally defined as a narrow tolerance margin around $\mathbf{M}$ (Eq. (13)), is redefined here. In the high-tolerance Ball Loss, we set $\mathbf{M}'$ to 1 at every voxel inside organ o except the N_i voxels in $\mathbf{M}$ (Eq. (24)). When the Dice and cross-entropy losses are applied as in the standard Ball Loss (Eq. (14)), the multiplication by $(1-\mathbf{M}')$ now zeros the gradient at every non-located voxel inside the organ. The model is therefore supervised only at the N_i most probable voxels (where $\mathbf{M}=1$, gradient pushes probabilities toward 1) and outside the organ (where both masks are 0, gradient pushes probabilities toward 0); everywhere else, the model is free to segment larger or additional tumors without penalty. This is shown in Eq. (24), where $\mathbf{O}$ is the pre-saved organ segmentation mask for organ o.

$$\mathbf{M}'=\mathbf{O}-\mathbf{M}\tag{24}$$

We then apply the same Dice and cross-entropy losses as in the standard Ball Loss (Eq. (14) to (16), copied below), using $\mathbf{M}$ as the target. The multiplication by $(1-\mathbf{M}')$ zeros the loss gradient at every voxel in $\mathbf{M}'$, leaving them not penalized.

$$\begin{aligned}&\mathcal{L}'_{\text{ball},o}=\mathrm{CE}((1-\mathbf{M}')\odot\mathbf{T}^o,\mathbf{M})+\\&\mathrm{Dice}((1-\mathbf{M}')\odot\mathbf{T}^o,\mathbf{M})\end{aligned}\tag{25}$$

$$\mathcal{L}_{\text{ball},o}=\begin{cases}\mathcal{L}'_{\text{ball},o}, & \text{if tumors are reported in } o\\ \mathrm{CE}(\mathbf{T}^o,\mathbf{0}), & \text{otherwise}\end{cases}\tag{26}$$

Table 6
95% confidence intervals (lower–upper, percentage points) for the detection metrics and standard deviations for DSC/NSD in Table 1. The 95% confidence intervals were estimated by nonparametric bootstrap over the test cases (1000 resamples; 2.5th and 97.5th percentiles). Statistical tests were performed to compare F1-Score and AUC (paired permutation test for F1-Score and DeLong's test for AUC). Orange highlights statistically significant gains ($p < 0.05$).

Train paradigm	Pancreas tumor												Kidney tumor					
	JHH-Test								UCSF-Test				UCSF-Test					
	mask	rep.	dsc	nsd	F1	AUC	Se	Sp	F1	AUC	Se	Sp	mask	rep.	F1	AUC	Se	Sp
few training masks (50)																		
standard segmentation	50	0	47	46	50–73	51–73	48–75	50–76	43–58	57–70	39–55	72–82	50	0	58–69	63–73	61–75	56–69
R-Super (ours)	50	2.2K	47	47	55–77	65–83	56–80	51–78	55–67	65–75	66–80	52–65	50	2.7K	60–70	60–71	66–80	48–62
medium / many training masks (344 / 1.7K)																		
CLIP-Like	344	2.2K	46	46	75–90	77–92	82–98	58–84	53–66	68–79	52–67	70–81	1.7K	2.7K	59–70	66–76	68–81	40–55
Multi-task learning	344	2.2K	44	44	62–81	70–88	66–89	46–74	46–59	56–68	57–73	43–56	1.7K	2.7K	58–70	65–76	61–76	56–70
Report-G Pseudo-labels	344	2.2K	44	43	72–88	68–87	69–90	65–88	62–75	78–87	57–72	81–90	1.7K	2.7K	66–75	69–79	74–86	53–66
Models Genesis	344	0	46	45	68–87	71–89	65–90	62–86	52–65	67–78	58–74	57–69	1.7K	0	61–72	68–78	71–83	47–60
nnU-Net	344	0	47	45	64–83	68–86	63–86	61–85	62–74	74–84	67–81	68–80	1.7K	0	67–77	70–80	82–92	45–60
standard segmentation	344	0	45	42	74–90	75–92	66–89	79–96	60–73	72–83	54–69	81–90	1.7K	0	66–75	68–78	58–72	61–74
R-Super (ours)	344	2.2K	41	38	85–96	85–97	87–100	78–96	78–87	87–94	76–89	85–93	1.7K	2.7K	69–79	73–82	74–85	64–77

$$L_{\text{ball}} = \sum_{o \in \mathcal{O}} L_{\text{ball},o} \tag{27}$$

The result is a loss that maximizes tumor probabilities in a small number of voxels (the most likely voxels for a 5 mm tumor), minimizes probabilities outside the organ o as usual, and applies no loss or gradient to all other voxels inside the organ o. The model is therefore encouraged to segment at least one small tumor, but it is free to segment a larger tumor or multiple tumors, without penalty.

Appendix B. Training details

We trained all segmentation models on 3D CT patches of $128 \times 128 \times 128$ voxels at an isotropic spacing of 1 mm. For CT-Report pairs, each patch was centered on a target organ (selected with 80% probability among the organs reported as containing tumors) and was sized to fully contain that organ. Without this constraint, a tumor mentioned in the report could fall outside the patch, and the report-based losses would push the model toward a tumor that is not visible in the input.

Optimization followed the MedFormer defaults (Gao et al., 2022): AdamW with weight decay 5×10^{-2}, gradient norm clipping at 1, batch size 2, and 100 epochs of 1000 batches each. The learning rate started at 1×10^{-4} with a 5-epoch warmup followed by polynomial decay. CT intensities were clipped to $[-991, 500]$ HU and then normalized. Data augmentation comprised rotation, brightness, gamma, contrast, Gaussian blur, and Gaussian noise, applied per sample as in MedFormer. R-Super-specific hyper-parameters are the loss weights, fixed across all experiments at 1 for the segmentation losses (Dice and cross-entropy on CT-Mask pairs), and 0.1 for the Volume and Ball Losses. These weights were tuned once on the UCSF-Train validation set and reused throughout, so no per-experiment tuning is required. When CT-Mask and CT-Report pairs were combined, CT-Mask pairs were super-sampled to 50% of each epoch's training samples.

All baselines compared to R-Super (standard segmentation, MTL, CLIP, Models Genesis, RG pseudo-labels) used the same MedFormer backbone and hyper-parameters as R-Super, so that differences in performance reflect the training methodology rather than the architecture. MTL additionally has a classification head attached to the MedFormer encoder output. Standard segmentation was trained on the CT-Mask pairs; MTL was trained on both CT-Mask and CT-Report pairs; CLIP was first pre-trained on the CT-Report pairs with a CLIP loss and then fine-tuned on the CT-Mask pairs; Models Genesis was first pre-trained on our CT scans with a masked-autoencoder loss and then fine-tuned on the CT-Mask pairs; report-guided pseudo-labels used the standard segmentation model to create pseudo-masks, masks were refined following (Bosma et al., 2023), and a new standard segmentation model (here, also MedFormer) was trained on these pseudo-masks plus the original tumor masks (following the same hyper-parameters as the other models).

The nnU-Net (Isensee et al., 2021) is the only method with a different architecture and training scheme. As a self-configuring framework intended to be used out of the box, it was trained on CT-Mask pairs using its standard procedure, automatic hyper-parameters, and ResEncL architecture. The only non-default setting was the isotropic voxel spacing of 1 mm, matching the spacing used by all other models.

Appendix C. Result variability

Tables 6 to 10 report the variability associated with the results in the corresponding tables of the main paper. For the segmentation metrics (DSC and NSD), we report the standard deviation. These standard deviations are large because tumor segmentation performance varies substantially across patients. Tumors in different patients vary considerably in size, shape, and texture, leading to differences in segmentation difficulty and, consequently, varying DSC and NSD scores. Although organs exhibit less variation than tumors, organ segmentation also yields large standard deviations in DSC and NSD (Bassi et al., 2024b). For the tumor detection metrics (sensitivity, specificity, F1 score, and AUC), we report 95% confidence intervals, calculated using 1000 bootstrap samples. Wide confidence intervals or standard deviations for individual results do not necessarily mean that the differences between methods are not statistically significant. Metrics that are significantly higher (p<0.05) are highlighted in orange in Tables 6 to 9.

Appendix D. LLM prompts

We used the prompt below to extract tumor information from radiology reports. The prompt was iteratively created in a collaboration between computer scientists and radiologists:

Instructions: The radiology report below possibly mentions one or more focal lesions (e.g., tumor, mass, nodule, cyst). Read it carefully, paying special attention to the findings, clinical history, and impressions sections (if available). Your task is to list the types, certainty of lesion type, sizes, organ, locations and attenuation of all lesions in the report. Fill out the template below, using one line per lesion (you may add or remove lines from the template): lesion 1: type = _; certainty = _; size = _; organ = _; location = _; attenuation = _; lesion 2: type = _; certainty = _; size = _; organ = _; location = _; attenuation = _; ... If you are absolutely sure the report mentions no lesion, do not use the template. Instead, reply with: 'No lesions mentioned.' and justify why you are sure the report mentions no lesion. Consider the following instructions: A - What is a Lesion: A lesion is any focal abnormality, including masses, cysts, or areas of altered density (hyperdense, hypodense, or isodense). You must list both benign and malignant lesions. Include lesions that are confirmed as well as those that are only suspicious (in this case, use 'certainty = low' in the filled template). Common terms that indicate a lesion: metastasis, nodule, soft tissue, nodular thickening, tumor, lesion, mass, cyst, pseudo-cyst, neoplasm, cancer, index lesion, oncologic finding, adenoma, carcinoma, growth, abnormal thickening, focus, LI-RADS lesion, hyperdensity, hypodensity, and isodensity. Terms

Table 7
95% confidence intervals (lower–upper, percentage points) for the detection metrics and standard deviations for DSC/NSD in Table 2 (size-stratified). The 95% confidence intervals were estimated by nonparametric bootstrap over the test cases (1000 resamples; 2.5th and 97.5th percentiles). Statistical tests were performed to compare F1-Score and AUC (paired permutation test for F1-Score and DeLong's test for AUC). Orange highlights statistically significant gains ($p < 0.05$).

Train paradigm	Pancreas tumor												Kidney tumor					
	JHH-Test								UCSF-Test				UCSF-Test					
	mask	rep.	dsc	nsd	F1	AUC	Se	Sp	F1	AUC	Se	Sp	mask	rep.	F1	AUC	Se	Sp
small tumors (diameter ≤ 2 cm)																		
few training masks (50)																		
standard segmentation	50	0	13	29	18–56	43–78	27–78	48–76	32–50	51–66	31–51	72–83	50	0	34–51	52–67	42–64	56–70
R-Super (ours)	50	2.2K	24	32	15–52	51–83	21–73	50–77	46–61	65–78	66–84	52–65	50	2.7K	38–54	53–69	54–75	48–62
medium / many training masks (344 / 1.7K)																		
CLIP-Like	344	2.2K	19	32	52–82	82–97	100–100	58–83	41–58	67–79	45–64	69–81	1.7K	2.7K	32–48	58–73	46–68	41–54
Multi-task learning	344	2.2K	20	31	35–70	69–94	65–100	46–74	34–50	54–68	49–71	43–56	1.7K	2.7K	37–54	57–73	46–68	56–70
Report-G Pseudo-labels	344	2.2K	21	36	40–77	62–91	50–94	68–90	54–71	76–87	53–72	81–90	1.7K	2.7K	41–57	64–77	56–76	53–66
Models Genesis	344	0	15	27	40–78	74–94	58–100	63–87	40–55	63–76	50–70	57–70	1.7K	0	34–52	58–72	48–71	47–61
nnU-Net	344	0	26	30	36–74	61–91	50–94	62–86	53–68	72–84	62–80	68–79	1.7K	0	45–61	64–78	72–90	46–59
standard segmentation	344	0	20	31	42–81	69–95	40–92	78–96	50–67	70–83	48–67	80–89	1.7K	0	29–48	62–76	33–56	60–74
R-Super (ours)	344	2.2K	26	31	62–93	80–96	79–100	78–96	68–82	86–94	68–85	85–93	1.7K	2.7K	48–64	72–84	59–79	62–76
large tumors (diameter > 2 cm)																		
few training masks (50)																		
standard segmentation	50	0	6	10	45–71	48–73	49–80	49–75	23–48	67–87	42–78	72–83	50	0	47–65	68–82	71–91	56–69
R-Super (ours)	50	2.2K	15	24	52–78	68–87	62–90	51–77	21–41	65–85	62–93	52–65	50	2.7K	45–61	65–78	77–93	48–62
medium / many training masks (344 / 1.7K)																		
CLIP-Like	344	2.2K	19	28	64–85	73–91	72–97	58–84	30–54	73–93	67–96	69–80	1.7K	2.7K	46–62	71–83	89–100	41–54
Multi-task learning	344	2.2K	22	32	49–75	66–88	58–88	47–73	19–37	59–80	68–96	44–56	1.7K	2.7K	46–62	70–83	68–87	56–70
Report-G Pseudo-labels	344	2.2K	22	34	67–87	67–89	70–94	67–90	42–67	81–97	68–95	81–90	1.7K	2.7K	54–69	69–81	92–100	53–67
Models Genesis	344	0	18	26	58–82	68–87	62–90	63–87	26–46	72–92	71–97	57–70	1.7K	0	50–65	74–86	92–100	47–61
nnU-Net	344	0	26	30	58–81	68–88	61–90	62–85	33–56	84–94	76–100	68–79	1.7K	0	48–63	70–82	87–99	46–59
standard segmentation	344	0	22	31	71–91	73–93	68–94	78–96	42–67	77–96	71–97	80–90	1.7K	0	55–71	68–81	84–97	61–74
R-Super (ours)	344	2.2K	27	34	79–96	86–98	86–100	78–96	56–79	89–97	89–100	85–93	1.7K	2.7K	59–74	69–81	92–100	62–75

Table 8
95% confidence intervals (lower–upper, percentage points) for the detection metrics and standard deviations for DSC/NSD in Table 3. The 95% confidence intervals were estimated by nonparametric bootstrap over the test cases (1000 resamples; 2.5th and 97.5th percentiles). Statistical tests were performed to compare R-Super and standard segmentation for F1-Score and AUC (paired permutation test for F1-Score and DeLong's test for AUC); orange highlights statistically significant gains ($p < 0.05$).

Train paradigm	Training		JHH-Small						JHH-Large						Merlin-Test			
	mask	rep.	dsc	nsd	F1	AUC	Se	Sp	dsc	nsd	F1	AUC	Se	Sp	F1	AUC	Se	Sp
nnU-Net	926	0	26	30	79–88	87–94	74–87	82–93	26	30	76–92	90–99	78–97	90–98	64–75	71–80	61–75	67–79
Multi-task learning	926	1,848	25	32	87–95	91–98	85–97	86–97	24	24	93–100	96–100	93–100	90–100	64–75	71–79	56–70	77–88
standard segmentation	926	0	24	31	86–95	90–97	78–93	94–100	22	21	96–100	99–100	93–100	100–100	61–73	69–78	52–67	77–87
R-Super (ours)	926	1,848	24	29	92–98	94–99	87–98	94–100	23	22	95–100	100–100	100–100	90–100	73–82	79–87	70–83	74–85

Table 9
95% confidence intervals (lower–upper, percentage points) for the detection metrics in Table 4. The 95% confidence intervals were estimated by non-parametric bootstrap over the test cases (1000 resamples; 2.5th and 97.5th percentiles). Statistical tests compare R-Super and standard segmentation (paired permutation test for F1-Score, DeLong's test for AUC); orange highlights statistically significant gains ($p < 0.05$).

Train paradigm	Training		Swiss			
	mask	rep.	F1	AUC	Se	Sp
Swiss-Test						
nnU-Net	3,488	0	74–81	70–78	70–78	65–77
Multi-task learning	3,488	28,295	70–77	72–79	59–68	80–89
standard segmentation	3,488	0	76–84	79–85	67–77	80–91
R-Super (ours)	3,488	28,295	83–88	82–88	78–85	81–90
UCSF-Test-Non-Contrast						
nnU-Net	3,488	0	40–62	58–76	52–80	66–80
Multi-task learning	3,488	28,295	5–33	50–60	2–20	98–100
standard segmentation	3,488	0	47–74	67–82	36–65	93–99
R-Super (ours)	3,488	28,295	63–84	88–96	55–80	93–99

*that are not a focal lesion: diverticulum (unless it is suspicious for malignancy); renal/gallbladder stones (e.g., cholelithiasis); hyper/hypoenhancing fluid collection not caused by a cyst or mass (e.g., caused by abscess, or postoperative changes); if the patient has cancer history but the tumor was surgically removed (e.g., whipple procedure for pancreatic tumors) and there is no current evidence of the tumor, do not list it; B - Size: 1- When to use numbers: Reports can write the size of the lesion in 1D, 2D, or 3D measurements, and you should use the same standards used in the report. Write 1D measurements as: 15 mm; 2D measurements as: 15 x 10 mm; and 3D measurements as: 40 x 30 x 30 mm. You may use either cm or mm, but you MUST WRITE in each line of the filled template the unit you are using (cm or mm). If a report does not specify the unit, assume it is mm. Say size = U if the report does not specify the size of a lesion. 2- When to use 'size = tiny' or 'size = massive': If a report describes a lesion without numeric diameters, but describes the lesion with adjectives like tiny, small, large, or massive, you **must** include one entry with*

Table 10
Uncertainty for Table 5: 95% confidence intervals (lower–upper, percentage points) for Se/Sp/F1/AUC and standard deviations for DSC/NSD. The 95% confidence intervals were estimated by nonparametric bootstrap over the test cases (1000 resamples; 2.5th and 97.5th percentiles).

Train paradigm	Pancreas tumor												Kidney tumor					
	JHH-Test								UCSF-Test				UCSF-Test					
	mask	rep.	dsc	nsd	F1	AUC	Se	Sp	F1	AUC	Se	Sp	mask	rep.	F1	AUC	Se	Sp
R-Super (ours)	344	2.2 K	41	38	85–96	85–97	87–100	78–96	78–87	87–94	76–89	85–93	1.7 K	2.7 K	69–79	73–82	74–85	64–77
loss functions																		
Volume Loss only	344	2.2 K	41	40	88–98	94–100	81–98	94–100	70–81	85–93	73–86	77–86	1.7 K	2.7 K	70–80	69–79	81–91	56–70
Ball Loss only	344	2.2 K	43	42	74–91	86–96	66–90	81–98	64–77	77–87	56–72	86–94	1.7 K	2.7 K	69–78	69–80	82–92	51–65
LLM size																		
small LLM (4B)	344	2.2 K	40	34	73–90	75–92	63–86	84–98	64–76	74–85	66–81	74–84	1.7 K	2.7 K	66–76	71–80	85–94	41–55
tiny LLM (0.6B)	344	2.2 K	44	42	63–83	72–88	57–83	68–90	55–70	68–78	45–63	86–94	1.7 K	2.7 K	68–77	71–81	81–91	48–62
report information removed																		
no pancreas subsegment	344	2.2 K	43	37	76–92	83–96	68–90	84–98	65–77	76–85	65–81	76–86	1.7 K	2.7 K	67–77	70–80	81–92	47–60
no tumor size	344	2.2 K	45	42	76–93	81–94	63–88	93–100	60–73	72–81	56–72	79–88	1.7 K	2.7 K	68–78	71–81	82–93	48–62
organ mask importance																		
50 CT organ segmenter	344	2.2 K	39	31	89–99	93–100	89–100	86–100	70–81	81–89	80–91	71–82	1.7 K	2.7 K	67–77	70–80	78–89	50–64
no organ mask	344	2.2 K	42	37	80–94	87–97	67–91	93–100	63–75	76–85	60–76	78–88	1.7 K	2.7 K	67–78	71–81	79–90	51–65
injected LLM errors (or report errors)																		
+5% LLM errors	344	2.2 K	39	32	86–97	88–98	81–98	90–100	66–78	80–89	65–80	78–88	1.7 K	2.7 K	69–78	72–81	86–94	44–58
+10% LLM errors	344	2.2 K	41	34	82–95	90–98	78–96	83–98	72–83	83–91	71–85	83–91	1.7 K	2.7 K	68–78	72–81	77–88	56–70
+20% LLM errors	344	2.2 K	41	34	82–95	83–96	76–94	87–100	65–77	77–86	68–83	73–84	1.7 K	2.7 K	67–77	71–81	81–92	46–60
100% LLM errors	344	2.2 K	45	42	57–80	64–84	45–73	80–98	50–62	62–73	65–80	44–57	1.7 K	2.7 K	60–71	64–75	70–83	43–57

*'size = tiny' or 'size = massive' for that lesion. However, always prefer using the lesion diameter, if provided. If the report does not provide the diameter nor any size-related adjective, say size = U. 3- When to use 'size = multiple': If the report mentions multiple lesions in an organ **but does not give an exact count**, you **must** include one entry with 'size = multiple' for that organ. Only use 'size = multiple' when you cannot determine the number; otherwise, create individual entries for each described lesion. Handling reports with unknown lesion counts AND some described lesions: If the report says there are multiple/innumerable/many lesions in an organ **and** then describes a few of them (e.g., gives their size), your output should include: One row describing each **explicitly described** lesion. One additional row with 'size = multiple' to represent the unspecified lesions. Example: Report: A 2 cm metastasis in liver sub-segment 3, and multiple other small metastases in the liver. Your output: lesion 1: type = metastasis; certainty = high; size = 2 cm; organ = liver; location = segment 3; attenuation = U; lesion 2: type = metastasis; certainty = high; size = multiple; organ = liver; location = U; attenuation = U; Key rules: 1. Always include individual entries for every lesion that is specifically described in the report. 2. Whenever the report indicates multiple unspecified lesions in an organ, also include one entry with 'size = multiple' for the organ. C - Organ: Organ is the organ where the lesion is located. Use standard organ names, like: liver, pancreas, kidney, spleen, colon, pelvis, adrenal gland, bladder, gallbladder, breast, stomach, lung, esophagus, uterus, bone, prostate, and duodenum. Do not consider the 'GI-Tract' as an organ. Instead, try to localize the tumor in one of these GI-tract organs: esophagus, stomach, duodenum, small intestines or colon. You can consider 'organ = colon' for rectal lesions. You can consider 'organ = esophagus' for lesions in the esophagus-gastric junction. D - Location: For location, check if the report specifies the sub-segment or part of the organ where the lesion is. If it specifies: for liver, choose location as segment 1/2/.../8. For the pancreas choose the pancreas head/body/neck/tail/uncinate process. For the kidney: left kidney/right kidney. For other organs, just check if the report mentions some type of organ region or sub-segment. Say location = U if the report does not specify the intra-organ location of a lesion. If a single lesion is in more than one location, you can say both. E.g., location = liver segment 4/5. E - Type: If the report provides lesion type, inform it. Otherwise, say type = U if the report does not specify the type of the lesion. Follow these rules: 1- If the lesion type is not specified in findings, you may deduce it from the clinical history or impressions sections. 2- Cyst, or cystic lesion, is a common type of lesion. For any cysts, you say type = cyst. 3- Assign type = metastasis if the lesion is described as a metastasis (or implant) originating from a cancer in another organ, unless the report explicitly states otherwise. Determine certainty based on the level of confidence expressed in the report. 4- If the report does not mention the lesion type, but you read that the patient has cancer in the organ where the lesion is, or has a history of malignant lesions in the organ, or the lesion is growing, or it is 'suspicious for malignancy', say type = malignant. However, do not say type = malignant if a more specific lesion type is given (like PDAC and PNET in pancreas, RCC in kidney, HCC in liver, ...). 4- If the report does not mention the lesion type, but it explicitly indicates that the lesion is likely benign (or mentions it may be one of several benign lesion types), say type = benign. However, do not say type = benign if a more specific lesion type is given (like cyst or polyp). 5- Try reporting types using their most standard name, followed by their acronym. For example, if the report mentions 'adenocarcinoma in the pancreas' or 'PDAC', say type = Pancreatic Ductal Adenocarcinoma (PDAC). F - Certainty: Certainty of the lesion type, according to the report. If a report mentions a lesion type in the findings, history, or impressions, without demonstrating uncertainty, say certainty = certain. If the report expresses strong confidence in lesion type, say certainty = high. If the report mentions a lesion type but expresses significant uncertainty about it, say certainty = low. If the report does not mention the lesion type, say certainty = U. G - Attenuation: For each lesion, inform the attenuation if the report mentions it. You should choose one of these options: hyperenhancing, hypoenhancing, isoenhancing, heterogeneously enhancing, or U (unknown). The report may use synonyms like hypermetabolic and hypoattenuating, but you must only answer me hyperenhancing, hypoenhancing, isoenhancing, heterogeneously enhancing or U (unknown). H - Justification: Besides filling the template, justify your answer, carefully mentioning each section of the report if present: history, findings, and impressions. Explain from which sentences you got each size, location, and type. Some reports may refer to past measurements (using words like previously, before, or giving dates). Ignore previous measurements. Provide me with a syntactic analysis of the report sentences mentioning lesion sizes. In this analysis, explain which measurement refers to which lesion, if the measurement is current or past, and if the corresponding lesion is malignant or benign. I will provide an example of a report and a correct answer for a %(organ)s lesion: Example report: %(example_report)s Example answer: %(example_answer)s End of the example.*

Data availability

Code, dataset, and models are available at https://github.com/MrGiovanni/R-Super.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., et al., 2021. The medical segmentation decathlon. arXiv preprint arXiv:2106.05735.

Bassi, P.R., Dertkigil, S.S., Cavalli, A., 2024a. Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization. Nat. Commun. 15 (1), 291.

Bassi, P.R., Li, W., Chen, J., Zhu, Z., Lin, T., Decherchi, S., Cavalli, A., Wang, K., Yang, Y., Yuille, A.L., Zhou, Z., 2025a. Learning segmentation from radiology reports. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 305–315, URL https://github.com/MrGiovanni/R-Super.

Bassi, P.R., Li, W., Tang, Y., Isensee, F., Wang, Z., Chen, J., Chou, Y.-C., Kirchhoff, Y., Rokuss, M., Huang, Z., Ye, J., He, J., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K.H., Jaeger, P., Ye, Y., Xie, Y., Zhang, J., Chen, Z., Xia, Y., Xing, Z., Zhu, L., Sadegheih, Y., Bozorgpour, A., Kumari, P., Azad, R., Merhof, D., Shi, P., Ma, T., Du, Y., Bai, F., Huang, T., Zhao, B., Wang, H., Li, X., Gu, H., Dong, H., Yang, J., Mazurowski, M.A., Gupta, S., Wu, L., Zhuang, J., Chen, H., Roth, H., Xu, D., Blaschko, M.B., Decherchi, S., Cavalli, A., Yuille, A.L., Zhou, Z., 2024b. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets Benchmarks Track 37, 15184–15201, URL https://github.com/MrGiovanni/Touchstone.

Bassi, P.R., Yavuz, M.C., Hamamci, I.E., Er, S., Chen, X., Li, W., Menze, B., Decherchi, S., Cavalli, A., Wang, K., Yang, Y., Yuille, A., Zhou, Z., 2025b. Radgpt: Constructing 3d image-text tumor datasets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23720–23730, URL https://github.com/MrGiovanni/RadGPT.

Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.-B., Reis, E., Truyts, C., et al., 2024. Merlin: A vision language foundation model for 3d computed tomography. Res. Sq. rs–3.

Bosma, J.S., Saha, A., Hosseinzadeh, M., Slootweg, I., de Rooij, M., Huisman, H., 2023. Semisupervised learning with report-guided pseudo labels for deep learning–based prostate cancer detection using biparametric MRI. Radiol.: Artif. Intell. 5 (5), e230031.

Can, Y.B., Chaitanya, K., Mustafa, B., Koch, L.M., Konukoglu, E., Baumgartner, C.F., 2018. Learning to segment medical images with scribble-supervision alone. In: Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R., Bradley, A., Papa, J.P., Belagiannis, V., Nascimento, J.C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., Madabhushi, A. (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer International Publishing, Cham, pp. 236–244.

Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., Tang, W., Jin, G., Jiang, H., Fang, X., et al., 2023. Large-scale pancreatic cancer detection via non-contrast CT and deep learning. Nature Med. 29 (12), 3033–3043.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), In: Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., pp. 12546–12558, URL https://proceedings.neurips.cc/paper_files/paper/2020/file/949686ecef4ee20a62d16b4a2d7ccca3-Paper.pdf.

Chen, E.Z., Dong, X., Li, X., Jiang, H., Rong, R., Wu, J., 2019. Lesion attributes segmentation for melanoma detection with multi-task u-net. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. ISBI 2019, IEEE, pp. 485–488.

Chen, Q., Zhou, X., Liu, C., Chen, H., Li, W., Jiang, Z., Huang, Z., Zhao, Y., Yu, D., He, J., Zheng, Y., Shao, L., Yuille, A., Zhou, Z., 2025. Scaling tumor segmentation: Best lessons from real and synthetic data. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 24001–24013, URL https://github.com/BodyMaps/AbdomenAtlas2.0.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al., 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al., 2009. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur. J. Cancer 45 (2), 228–247.

Gao, Y., Zhou, M., Liu, D., Yan, Z., Zhang, S., Metaxas, D.N., 2022. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. arXiv preprint arXiv:2203.00131.

Gobara, A., Yoshizako, T., Yoshida, R., Nakamura, M., Shiina, H., Kitagaki, H., 2019. T1a renal cell carcinoma on unenhanced CT: Analysis of detectability and imaging features. Acta Radiol. Open 8 (5), 2058460119849706. http://dx.doi.org/10.1177/2058460119849706.

Haghighi, F., Hosseinzadeh Taher, M.R., Zhou, Z., Gotway, M.B., Liang, J., 2020. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 137–147, URL https://github.com/fhaghighi/SemanticGenesis.

Hamamci, I.E., Er, S., Menze, B., 2024. Ct2rep: Automated radiology report generation for 3d medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 476–486.

Hoogenboom, S.A., Engels, M.M., Chuprin, A.V., van Hooft, J.E., LeGout, J.D., Wallace, M.B., Bolan, C.W., 2022. Prevalence, features, and explanations of missed and misinterpreted pancreatic cancer on imaging: a matched case–control study. Abdom. Radiol. 47 (12), 4160–4172.

Hooper, S., Chen, M., Saab, K., Bhatia, K., Langlotz, C., Ré, C., 2023. A case for reframing automated medical image classification as segmentation. Adv. Neural Inf. Process. Syst. 36, 55415–55441.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211.

Kolesnikov, A., Lampert, C.H., 2016. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: European Conference on Computer Vision. Springer, pp. 695–711.

Landman, B., Xu, Z., Eugenio Igelsias, J., Styner, M., Robin Langerak, T., Klein, A., 2017. Multi-atlas labeling beyond the cranial vault-workshop and challenge.

Li, W., Bassi, P.R., Lin, T., Chou, Y.C., Zhou, X., Tang, Y., Isensee, F., Wang, K., Chen, Q., Xu, X., Ye, J., Zhu, Z., Decherchi, S., Cavalli, A., Yuille, A.L., Zhou, Z., 2025a. ScaleMAI: Accelerating the development of trusted datasets and ai models. arXiv preprint arXiv:2501.03410, URL https://github.com/MrGiovanni/ScaleMAI.

Li, W., Bassi, P., Zhou, X., Wang, K., Yang, Y., Yuille, A., Zhou, Z., 2026a. Cancerverse: Multicenter CT segmentation of 16 cancers for radiotherapy targets and oars. In: 2026 Joint AAPM| COMP Annual Meeting & Exhibition. AAPM.

Li, W., Bassi, P.R., Zhou, X., Wasserthal, J., Yuille, A.L., Zhou, Z., 2026b. RadThinking: A dataset for longitudinal clinical reasoning in radiology. arXiv preprint arXiv:2605.10761.

Li, W., Qu, C., Chen, X., Bassi, P.R., Shi, Y., Lai, Y., Yu, Q., Xue, H., Chen, Y., Lin, X., Tang, Y., Cao, Y., Han, H., Zhang, Z., Liu, J., Zhang, T., Ma, Y., Wang, J., Zhang, G., Yuille, A., Zhou, Z., 2024. AbdomenAtlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. Med. Image Anal. 103285, URL https://github.com/MrGiovanni/AbdomenAtlas.

Li, W., Zhou, X., Chen, Q., Lin, T., Bassi, P.R., Chen, X., Ye, C., Zhu, Z., Ding, K., Li, H., Wang, K., Yang, Y., Tang, Y., Xu, D., Yuille, A.L., Zhou, Z., 2025b. Pants: The pancreatic tumor segmentation dataset. In: Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track. URL https://github.com/MrGiovanni/PanTS.

Ma, J., Kim, S., Li, F., Baharoon, M., Asakereh, R., Lyu, H., Wang, B., 2024. Segment anything in medical images and videos: Benchmark and deployment. arXiv preprint arXiv:2408.03322.

Miller, A., Hoogstraten, B., Staquet, M., Winkler, A., 1981. Reporting results of cancer treatment. Cancer 47 (1), 207–214.

Park, S., Chu, L., Fishman, E., Yuille, A., Vogelstein, B., Kinzler, K., Horton, K., Hruban, R., Zinreich, E., Fouladi, D.F., et al., 2020. Annotated normal CT data of the abdomen for deep learning: Challenges and strategies for implementation. Diagn. Interv. Imaging 101 (1), 35–44.

Qu, C., Zhang, T., Qiao, H., Liu, J., Tang, Y., Yuille, A., Zhou, Z., 2023. AbdomenAtlas-8K: Annotating 8,000 abdominal CT volumes for multi-organ segmentation in three weeks. In: Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track. 21, URL https://github.com/MrGiovanni/AbdomenAtlas.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Rajchl, M., Lee, M.C.H., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., Rueckert, D., 2017. DeepCut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans. Med. Imaging 36 (2), 674–683. http://dx.doi.org/10.1109/TMI.2016.2621185.

Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al., 2025. Medgemma technical report. arXiv preprint arXiv:2507.05201.

Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740.

Wald, T., Ulrich, C., Lukyanenko, S., Goncharov, A., Paderno, A., Miller, M., Maerkisch, L., Jäger, P.F., Maier-Hein, K., 2024. Revisiting mae pre-training for 3d medical image segmentation. arXiv preprint arXiv:2410.23132.

Xia, Y., Yu, Q., Chu, L., Kawamoto, S., Park, S., Liu, F., Chen, J., Zhu, Z., Li, B., Zhou, Z., Lu, Y., Wang, Y., Shen, W., Xie, L., Zhou, Y., Wolfgang, C., Javed, A., Fouladi, D.F., Shayesteh, S., Graves, J., Blanco, A., Zinreich, E.S., Kinny-Köster, B., Kinzler, K., Hruban, R.H., Vogelstein, B., Yuille, A.L., Fishman, E.K., 2022. The FELIX project: Deep networks to detect pancreatic neoplasms. MedRxiv.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P.,

Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z., 2025. Qwen3 technical report. URL https://arxiv.org/abs/2505.09388, arXiv:2505.09388.

Zeman, R.K., Zeiberg, A., Hayes, W.S., Silverman, P.M., Cooper, C., Garra, B.S., 1996. Helical CT of renal masses: The value of delayed scans. Am. J. Roentgenol. 167 (3), 771–776. http://dx.doi.org/10.2214/AJR.167.3.8751698.

Zhang, Y., Li, H., Du, J., Qin, J., Wang, T., Chen, Y., Liu, B., Gao, W., Ma, G., Lei, B., 2021. 3D multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification. IEEE Trans. Med. Imaging 40 (6), 1618–1631.

Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J., 2021. Models genesis. Med. Image Anal. 67, 101840, URL https://github.com/MrGiovanni/ModelsGenesis.

Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019. Models genesis: Generic autodidactic models for 3d medical image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 384–393, URL https://github.com/MrGiovanni/ModelsGenesis.

Zhu, J., Li, Y., Hu, Y., Ma, K., Zhou, S.K., Zheng, Y., 2020. Rubik's cube+: A self-supervised feature learning framework for 3D medical image analysis. Med. Image Anal. 64, 101746.