Large-Scale Multi-Cancer Detection by Learning Segmentation from Reports

Pedro R. A. S. Bassi[1,2,3*†], Xinze Zhou[1†], Wenxuan Li[1,2,3†], Szymon Płotka[4†], Jakob Wasserthal[5], Jieneng Chen[1], Ibrahim E. Hamamci[6,7], Sezgin Er[6,8], Jakub Prządo[9], Zheren Zhu[10,11], Gorkem Durak[12], Wiktoria Romańczyk[13], Yavuz B. Taktak[14], Melih Akan[8], Yuhan Wang[15], Scott Ye[11], Qi Chen[1], Ashwin Kumar[16], Lizhou Wu[17], Guang Zhang[17], Bjoern Menze[8], Jarosław B. Ćwikła[18], Yuyin Zhou[15,19], Frank H. Miller[12], Yunhe Gao[16], Akshay S. Chaudhari[16], Curtis P. Langlotz[16], Ulas Bagci[12], Sergio Decherchi[20], Andrea Cavalli[20,21,22], Arkadiusz Sitek[2,3*], Kang Wang[11*], Yang Yang[11*], Alan L. Yuille[1*], Zongwei Zhou[1,23*]

[1]Johns Hopkins University, Baltimore, MD, USA.
[2]Harvard Medical School, Boston, MA, USA.
[3]Massachusetts General Hospital, Boston, MA, USA.
[4]Jagiellonian University, Kraków, Poland.
[5]University Hospital Basel, Basel, Switzerland.
[6]University of Zurich, Zurich, Switzerland.
[7]ETH AI Center, Zurich, Switzerland.
[8]Istanbul Medipol University, Istanbul, Turkey.
[9]Warmian-Masurian Cancer Center, Olsztyn, Poland.
[10]University of California, Berkeley, CA, USA.
[11]University of California, San Francisco, CA, USA.
[12]Northwestern University, Illinois, USA.
[13]Medical University of Białystok, Poland.
[14]Istanbul University, Istanbul, Turkey.
[15]University of California, Santa Cruz, CA, USA.
[16]Stanford University, Stanford, CA, USA.
[17]Shandong Provincial Qianfoshan Hospital, Shandong, China.
[18]University of Warmia and Mazury, Olsztyn, Poland.
[19]Google, San Francisco, CA, USA.
[20]Istituto Italiano di Tecnologia, Genoa, Italy.
[21]University of Bologna, Bologna, Italy.
[22]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.
[23]Johns Hopkins Medicine, Baltimore, MD, USA.

*Corresponding author(s). E-mail(s): psalvad2@jh.edu; asitek@mgh.harvard.edu; kang.wang@ucsf.edu; yang.yang4@ucsf.edu; alan.yuille@jhu.edu; zzhou82@jh.edu;
†These authors contributed equally to this work.

Abstract

More than 300 million computed tomography (CT) scans are performed worldwide each year, yet many early or incidental tumors in these scans remain undetected. Artificial intelligence (AI) could help: segmentation models can surpass radiologists and alternative AI models in detecting tumors, and they localize the tumors for radiologist verification. However, segmentation-based tumor detection has long been limited by the need for tumor masks: radiologist-drawn tumor outlines that are scarce, expensive, and entirely unavailable for many cancer types. In contrast, nearly every CT scan is accompanied by a radiology report with detailed tumor descriptions. Yet, these reports have not been used effectively to train tumor segmentation models. Here, we introduce R-Super, a framework that converts routine radiology and pathology reports into localized training signals for tumor segmentation. R-Super trains AI to segment tumors that match their descriptions in reports. Reports are only needed for training, not inference. We trained R-Super on 127,496 CT-Report pairs (42 million 2D images, USA) and evaluated it internally at UCSF ($\boldsymbol{N}$=2,301, USA) and externally at

Stanford (N=1,976, USA), Medipol (N=1,327, Turkey), and Basel (N=2,935, Switzerland). R-Super detects 7 tumor types for which public tumor masks are scarce or absent: spleen, gallbladder, prostate, bladder, uterus, esophagus, and adrenal tumors. Training R-Super on over 100,000 reports (no mask) outperformed mask-based segmentation models trained on 870 masks, demonstrating that large-scale report-based training surpasses smaller-scale mask-based training. Alternatively, by training R-Super on both these reports and masks together, cancer detection sensitivity increased by over +11% beyond mask-only training, and DSC by +14%. R-Super significantly surpassed 6 alternative training frameworks trained on the same dataset and 9 leading public AI models. R-Super significantly surpassed six radiologists in detecting six tumor types and matched them for uterus tumors in a reader study. On average, R-Super detected 56% more malignant tumors than the radiologists, at matched false positive rate. These results show that radiology reports are not merely clinical documentation, but a large-scale, underutilized source of localized supervision for training more accurate cancer detection AI. By effectively learning from reports, R-Super enables tumor segmentation models to scale beyond scarce radiologist-drawn tumor masks and advances automated and incidental cancer detection closer to clinical deployment. We release code, over 22,000 CT scans and reports, and the first public AI model to reach or surpass radiologist performance in detecting these tumor types on CT.

1 Main

Cancer is a leading cause of death worldwide [1, 2], and early detection is crucial: five-year survival often exceeds 90% when tumors are detected early but drops below 20% once they become advanced or metastatic [3]. Each year, over 300 million Computed Tomography (CT) scans are performed globally [4], to evaluate suspected cancer and for many other reasons. In both cases, this enormous imaging volume registers numerous early-stage tumors that often go undetected, even by expert radiologists. For instance, in a study of CT scans taken before pancreatic cancer diagnosis, about 50% of the tumors were missed by radiologists [5]. Artificial Intelligence (AI) can assist radiologists in incidental or early tumor detection [6–8]: AI can perceive subtle tumor signs imperceptible to humans, analyze images in 3D rather than slice by slice, and has been shown to surpass radiologists in detecting pancreas [7, 9] and stomach [8] tumors on CT.

State-of-the-art AI for tumor detection is typically formulated as semantic segmentation [7–11], in which tumors are detected and localized by outlining them directly on the medical image. Compared with classification models, segmentation models are generally more accurate and more robust across hospitals [12, 13]. Their output is also clinically interpretable, because tumor outlines allow radiologists to inspect and verify the AI decision. However, training segmentation models requires *tumor masks*: voxel-wise tumor outlines drawn by radiologists. Creating these masks is slow, costly, and not part of standard radiologist work. Manually outlining a tumor on a single 3D CT scan can take over 30 minutes. In Xia *et al.* [6], generating 3,125 pancreatic tumor masks required eight radiologists, five years, and millions of dollars. Semi-automated and interactive contouring tools can reduce this burden, but they still require labor-intensive radiologist supervision and annotations [14]. As a result, public CT datasets contain few tumor masks (Fig. 1), which are concentrated in few organs, such as the kidney, liver, lungs, and pancreas [15–18]. For many other organs, including the spleen, gallbladder, prostate, and bladder, no public tumor masks exist. For the uterus, esophagus, and adrenal glands, fewer than 60 public tumor masks exist [19, 20]. Consequently, public AI models are not accurate in detecting these seven tumor types, performing substantially below radiologists (Tab. D20, D56).

While tumor masks are scarce, nearly every CT scan is already accompanied by a radiology report. These reports routinely describe tumor number, size, attenuation, and approximate anatomical location, while pathology reports provide confirmation of malignancy. CT-Report pairs are naturally orders of magnitude more abundant than CT-Mask pairs: public datasets already contain over 24,000 CT-Report pairs [21, 22], whereas the largest public CT-Mask datasets rarely exceed 1,000 tumor masks [15–18]. Similarly, hospitals may contain millions of CT-Report pairs, but few or no tumor masks. Reports describe tumors across all organs, while public tumor masks are limited to a few organs. This motivates a central question: *Can reports supplement, or even substitute, tumor masks in training AI for tumor segmentation?*

Reports have been used to train AI models for cancer-related tasks. Reports are commonly used to train foundation vision-language models (VLMs), including Google's MedGemma [23], Alibaba's Lingshu [24], Microsoft's MedImageInsight [25], and Stanford's Merlin [22]. Alternatively, labels extracted from reports are often used to train classifiers. However, VLMs and classifiers treat reports primarily as non-localized supervision: training signals that directly inform the AI model whether a tumor is present, but not where it is localized within the CT scan space. As a result, classifiers and VLMs can detect or describe tumors, but cannot provide per-voxel tumor localization. This limits a radiologist's ability to verify these AI models' decisions. Furthermore, classifiers and VLMs often miss substantially more tumors than segmentation models [10, 12]. This suggests that localized supervision is important for accurate tumor detection: tumor masks inform segmentation models exactly where tumors are, helping them learn to detect tumors that may occupy less than 0.0001% of a 3D CT scan. This raises the question of whether, like tumor masks, radiology reports can provide the localized supervision.

Here, we propose R-Super (Report Supervision), a training framework that converts radiology and pathology reports into localized supervision to train tumor segmentation models on large-scale CT-Report datasets (Fig. 1). The challenge is that, unlike tumor masks, reports do not provide voxel-wise tumor locations. However, reports provide detailed descriptions of tumors, including tumor number, rough location, size, attenuation, and malignancy. R-Super transforms these descriptions into localized training signals through new loss functions, which teach segmentation models to segment tumors consistent with their report descriptions. Therefore, R-Super can learn tumor segmentation from reports only, without masks. When some masks are also available, R-Super can train with both CT-Report and CT-Mask pairs. Reports are used only during training, not inference, and the R-Super training framework can be applied to any segmentation architecture. Thus, R-Super directly addresses the main bottleneck in tumor segmentation: the scarcity of tumor masks.

To train R-Super, we created a new dataset, comprising 127,496 CT-Report pairs (42 million 2D slices), primary extracted from 27 years of archived data from the UCSF hospital and its affiliated institutions in the USA. We evaluated R-Super across four hospital cohorts in three countries: UCSF (N=2,301, USA), Stanford (N=1,976, USA), Medipol University Hospital (N=1,327, Turkey), and University Hospital Basel (N=2,935, Switzerland). We studied seven tumor types with few or no public CT tumor masks: spleen, gallbladder, prostate, bladder, uterus, esophagus, and adrenal tumors. *Even without tumor masks, R-Super substantially outperforms 9 leading public AI models in tumor detection and segmentation*, surpassing public VLMs by Google, Stanford, Alibaba, and Microsoft by at least +18% balanced accuracy in malignant tumor detection in external validation, and surpassing public segmentation models (ULS and FLARE) by at least +9% balanced accuracy and +31% DSC (Sec. 2.1). *R-Super shows reports can substitute or complement tumor masks:* training on over 100,000 CT-Report pairs surpassed training on 870 CT-Mask pairs (created by 31 radiologists). Training on these reports and masks together surpassed mask-only training by at least +11% malignant tumor detection sensitivity and +14% segmentation DSC (Sec. 2.2; Supp. D.2). *R-Super is more effective in learning from reports:* it surpassed 6 alternative training strategies also trained on our dataset by +5% to +12% detection sensitivity and +14% segmentation DSC (e.g., report-based CLIP, multi-task learning, and classification). *R-Super surpasses radiologist performance:* in a reader study, R-Super outperformed six radiologists in detecting six of seven tumor types, and matched them for uterus tumors. It detected 56% more malignant tumors than the radiologists on average, at matched specificity (Fig. 5). These results suggest that R-Super could help radiologists detect tumors that would otherwise be missed. R-Super also surpassed radiologist performances in the literature (Supp. B). To the best of our knowledge, R-Super is the first public AI model shown to surpass (or even reach) radiologist performance for detecting these tumor types on CT.

The central insight of this study is that medical reports are not merely clinical documentation or image-level labels, but an abundant source of localized supervision. Across other AI fields, such as NLP and computer vision, major advances like ChatGPT and CLIP came from shifting from small manually annotated datasets to massive datasets supervised by readily available text or captions [26, 27]. Our results (Fig. 3a) suggest that a similar transition is possible in cancer detection and segmentation: from training on hundreds of expensive CT-Mask pairs to training on hundreds of thousands of CT-Report pairs readily available in hospitals. By transforming reports into localized supervision, R-Super enables tumor segmentation models to scale beyond the longstanding limitation of scarce tumor masks. More broadly, R-Super offers a more interpretable and accurate alternative to classification and current VLMs for multi-tumor detection, enabling public AI models to surpass radiologist performance. This paradigm can advance opportunistic cancer detection by flagging probable cancer cases that would otherwise remain undetected on routine CT, enabling timely follow-up examinations. Together with the release of code, trained models, over 22,000 CT-Report pairs (chest, abdomen, and pelvis), and tumor masks for multiple tumor types unavailable in public segmentation datasets, this work provides an open-source foundation for broader automated and incidental cancer detection.

2 Results

We evaluated R-Super on four tasks: tumor detection, that is, assessing whether each organ contains tumors (benign + malignant tumor vs no-tumor); malignant tumor detection, that is, tumor detection stratified for malignant tumors (malignant vs no-tumor); malignancy identification, that is, detecting tumors and identifying whether they are malignant (malignant vs benign + no-tumor); and tumor localization, that is, determining whether detected tumors overlap with tumor masks drawn by radiologists.

Malignancy is confirmed with pathology reports in the UCSF test set. In the Stanford, Medipol, and Basel test sets, malignancy is confirmed from radiology reports, considering lesions described as malignant with high certainty or already diagnosed active cancer documented in the patient's clinical history. In the reader study test set, malignancy is confirmed using either pathology reports or documentation of already diagnosed active cancer in radiology reports (dataset details in Sec. 4.1).

R-Super produces tumor detection and malignancy identification outputs using a classification head applied to the tumor segmentation output (Supp. C.1). Tumor detection through segmentation improves accuracy (Fig. 3)

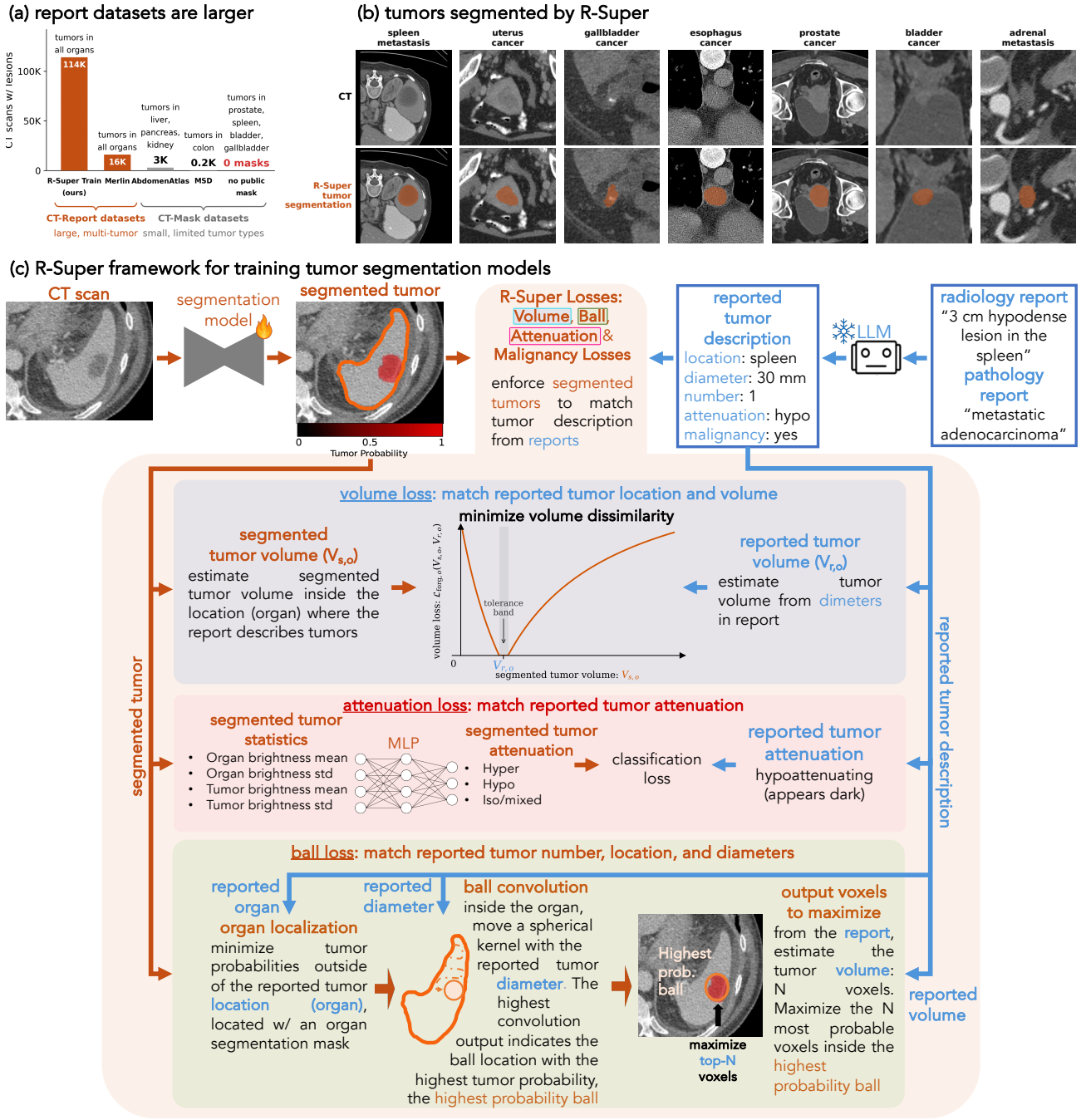

Fig. 1: **(a) R-Super learns from reports because reports are more abundant than tumor masks.** Our training set has 127K CT-Report pairs, 114K with tumors; Merlin [22] has 25K, 16K with tumors. The largest public CT-Mask datasets have up to 3K CT-Mask pairs with tumors. Many tumor types have no public masks. **(b) Examples of tumors segmented by R-Super** on CT scans from the external Medipol test set (Tab. 1). **(c) Overview of R-Super training framework.** First, an LLM (frozen) extracts tumor characteristics from radiology and pathology reports. Then, four new loss functions teach a tumor segmentation model to segment tumors that match these tumor characteristics. The **Volume Loss** (Sec. 4.3) enforces the volume of segmented tumors to match the report. From the tumor diameters in the radiology report, we estimate the total reported tumor volume in each organ o, $V_{r,o}$. We then sum the segmentation model output over all voxels inside organ o (located with a pre-saved AI-made organ mask) to estimate the segmented tumor volume, $V_{s,o}$. Finally, we encourage $V_{s,o}$ to match $V_{r,o}$ by minimizing a custom dissimilarity function (Eq. 8). The **Ball Loss** (Sec. 4.4) encourages segmented tumors to match the tumor number, rough locations, and diameters from the report. Each reported tumor is located by a Ball Convolution. It slides a fixed spherical kernel (a ball with the reported tumor diameter) over the segmentation output, summing the predicted tumor probabilities inside each sphere position. The convolution maximum is the most likely location for the reported tumor. Within this location, we maximize the top-N most probable voxels, where N is the tumor volume from report. These top-N voxels can have any shape within the reported tumor diameter. The Ball Loss locates and optimizes each tumor documented in the report, avoiding overlaps. The **Attenuation Loss** (Sec. 4.5) enforces segmented tumors to match the tumor attenuation (relative brightness) from the report. The **Malignancy Loss** (Sec. 4.6) distinguishes benign from malignant tumors using supervision from pathology reports if available.

and interpretability. For alternative segmentation models without classification heads (ULS, FLARE, and segmentation baselines), we computed tumor detection metrics by thresholding the segmentation output, following prior work [10].

Section 2.1 compares R-Super trained on reports only (no mask) to 9 public AI models in tumor detection. Section 2.2 shows how R-Super performance can be further scaled by training with CT-Mask and CT-Report pairs, and evaluates tumor localization. Section 2.3 compares R-Super to 6 alternative training methods also trained on our dataset. Finally, Sec. 2.4 compares R-Super to radiologists in a reader study. This paper includes three tasks, four test datasets, and comparisons to 15 alternative models. Given the large number of experiments, this Results section presents the main findings, while detailed results tables, additional analyses, and ablation studies are in Supp. D. Ablations show that all the R-Super loss functions positively contribute to the model performance (Supp. D.5). R-Super can train any segmentation architecture. Here, we chose MedFormer [28], for its success in previous benchmarks (Touchstone [29]). Data pre-processing and training details are in Supp. E.

2.1 Detect 7 tumor types without any tumor masks

We evaluated R-Super in the detection of 7 tumors types: spleen, gallbladder, prostate, bladder, uterus, esophagus, and adrenal gland tumors, 7 important tumor types underrepresented in public segmentation datasets. We compared R-Super with 9 leading public AI models, showing that learning tumor segmentation from reports allows R-Super to set a new state-of-the-art for open-source AI in tumor detection. Results are summarized in Fig. 2. Here, we trained R-Super on only CT-Report pairs (no mask). Medical AI often loses performance when tested on hospitals not seen during training, making evaluation on external (unseen) hospitals critical [6, 13]. Therefore, we evaluated R-Super on 3 external hospitals: Stanford Test (N=1,976, USA), Medipol (N=1,327, Turkey), and Basel (N=2,935, Switzerland); plus one internal: UCSF Test (N=2,301, USA). For the evaluation on Stanford Test, we trained R-Super on the UCSF Train set (N=116,514, USA). For evaluation on UCSF Test, Medipol, and Basel, we trained on the full R-Super train set (N=127,496, combining UCSF and Stanford data). Stanford data derives from the Merlin dataset [22], the largest public abdominal CT dataset, and, up to now, the only public dataset with an adequate number of the tumor types evaluated here. Details on all datasets are in Tab. 1 and A1. The Medipol and Basel test sets provide a particularly challenging test of generalization: these datasets include only images from Asia and Europe, but R-Super was trained only on North American images.

R-Super substantially surpassed 9 public AI models in tumor detection (Fig. 2 and Supp. D.6), including leading medical VLMs, namely Alibaba's Lingshu [24], Google's MedGemma [23], Microsoft's MedImageInsight [25], and Stanford's Merlin [22]. In malignant tumor detection (malignant vs no-tumor), R-Super significantly ($p < 0.05$) surpassed the best public AI model by +9% to +22% average balanced accuracy across all the four test sets (Fig. 2). Specifically in Medipol and Basel, R-Super outperformed all VLMs by +18% to +19% average balanced accuracy. These two datasets are important because they represent external validation for both R-Super and all VLMs (Merlin was trained at Stanford). In tumor detection (malignant + benign vs no-tumor), R-Super significantly ($p < 0.05$) surpassed the best public AI model by +10% to +19% average balanced accuracy (Supp. D.6). Among the public VLMs, zero-shot classification [30] performed substantially better than report generation. For instance, when evaluated with zero-shot classification, Merlin was the strongest public model (after R-Super) in 3 test sets. However, Merlin and the other VLMs missed almost all tumors when evaluated through report generation, with average sensitivity below 10% in every dataset (Supp. D.6). In contrast, R-Super achieved high sensitivity and specificity in all datasets. In malignant tumor detection, it achieved sensitivity and specificity of 87% (95% CI of 84-90%) and 87% (84-90%) in the UCSF test set, 84% (79-88%) and 91% (89-93%) in Stanford, 87% (82-91%) and 92% (89-94%) in Medipol, and 78% (73-83%) and 87 (83-90%) in Basel (detailed results in Fig. 2 and Supp. D.1). R-Super could even match or surpass human experts in a reader study (Sec. 2.4).

R-Super also surpassed public segmentation models by large margins. Our 9 baselines include multiple leading public segmentation models: the Universal Lesion Segmentation model (ULS) [31], the top two segmentation models in the FLARE 23 challenge [32], the top FLARE 24 model [33], and the FLARE 25 baseline model [34]. Unlike R-Super, these models were trained only with tumor masks, not reports, and their training datasets do not distinguish tumor type or malignancy. These public segmentation models underperformed in our study (Fig. 2), suggesting that the tumor types evaluated here are underrepresented in the FLARE and ULS datasets. This finding highlights how limited tumor mask availability remains a key bottleneck for open research on multi-tumor detection and localization. As an additional limitation, ULS was only trained on images cropped around tumors [31], missing no-tumor controls, which can limit tumor detection performance. Even so, ULS performed comparably to the FLARE models (Fig. 2). Overall, our results (Fig. 2, Supp. D.6) indicate that R-Super is, to our knowledge, the first public AI model capable of reliably detecting these 7 tumor types on CT.

Besides detecting tumors, R-Super identifies whether they are malignant. In malignancy identification (malignant vs. benign + no-tumor), R-Super achieved on average: 84% (81-86%) AUC on UCSF test (internal, USA), 85% (82-88%) on Stanford test (external, USA), 76% (71-79%) on Basel test (external, Switzerland), and 88% (85-90%) on Medipol test (external, Turkey). Results are summarized in Fig. 2 and detailed in

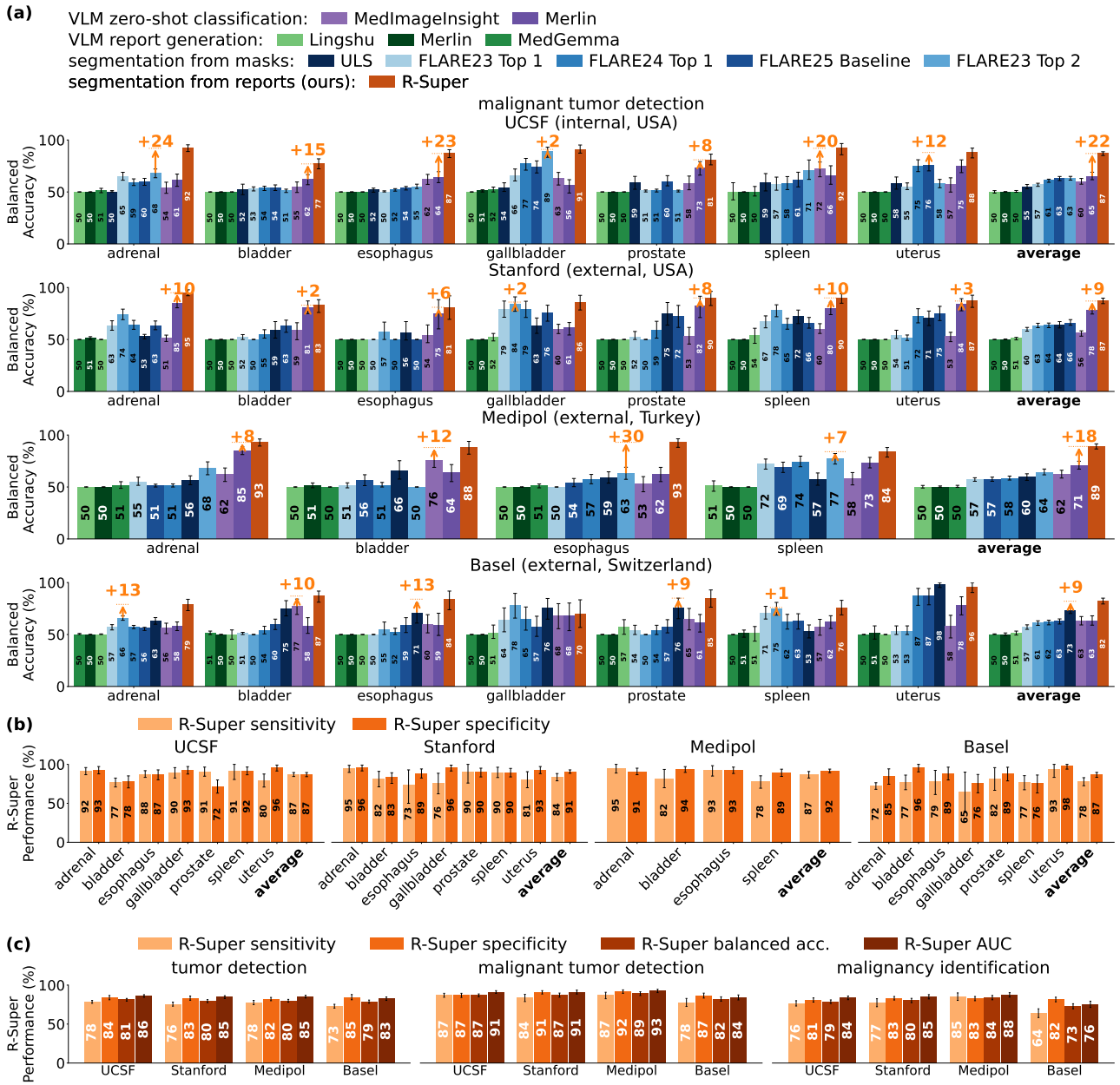

Fig. 2: **(a) R-Super, trained with reports only (no mask), surpasses leading public AI models in multi-tumor detection, by large margins.** R-Super surpassed the public AI models by +9% to +22% average balanced accuracy. In this figure, we evaluate the AI models for malignant tumor detection (malignant vs no-tumor). We skipped prostate, uterus, and gallbladder tumors in the Medipol test set, where the number of these malignant tumors was insufficient (Tab. A1). The plot shows balanced accuracy, at the best operating point for each model. The results of all models represent external evaluation on a hospital not seen in training, except for R-Super evaluated at UCSF, and Merlin evaluated at Stanford. Details on the VLM evaluation procedure are available in Supp. F. **(b) Sensitivity and specificity of R-Super** in malignant tumor detection (malignant vs no-tumor). **(c) Performance of R-Super in tumor detection, malignant tumor detection, and malignancy identification.** Detailed results for all models are available in Supp. D.6. Dataset details are in Tab. 1 and Tab. A1. Error bars represent 95% confidence intervals, calculated with bootstrapping (1,000 bootstrap samples).

Tab. D6. The public AI models were not evaluated for malignancy identification, because the public segmentation models were not trained for this task, and public VLMs had insufficient tumor detection performance (tumors must first be detected before malignancy can be assessed). Malignancy identification is a very challenging task even for humans, and the average performance of R-Super surpassed radiologists in this task (Sec. 2.4). However, AI-based tumor detection is arguably more critical than malignancy identification: a detected tumor of unknown malignancy can be characterized as malignant or not through follow-up exams, but an undetected tumor may go untreated.

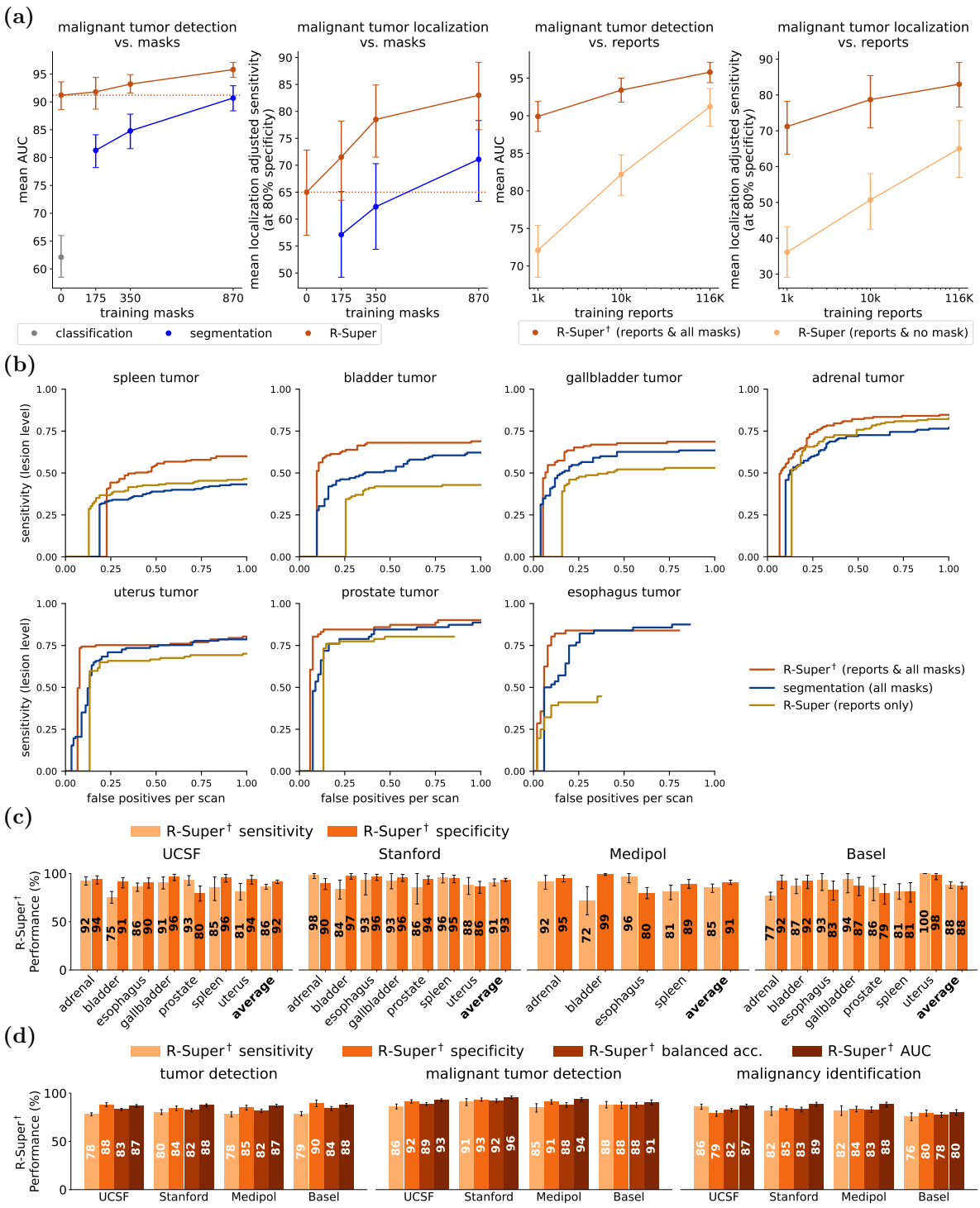

Fig. 3: **Training with many reports surpasses training with few masks; CT-Report plus CT-Mask training surpasses CT-Mask training. (a) Scaling: R-Super performance improves with more training reports or masks.** R-Super trained on reports alone (116,514) surpassed a segmentation model trained on 870 masks for tumor detection, and one trained on 350 masks for tumor localization. For any number of training masks, adding CT-Report pairs to training substantially improved performance. Two left plots: scaling training masks. Two right plots: scaling reports. Localization-adjusted sensitivity: percentage of scans where AI-generated masks overlap radiologist-drawn masks. The segmentation model (blue) was trained on CT-Mask pairs only. Models were trained on UCSF train, and externally validated on Stanford test (N=1,976 scans; N=602 masks used for localization evaluation; Tab. 1). **(b) R-Super trained on reports and masks achieves higher FROC (tumor-level localization) than mask-only segmentation training for all tumor types.** For adrenal and uterus tumors, R-Super surpasses mask-only even if trained on reports alone. DSC in Fig. D6 and D5. **(c) Sensitivity & specificity of R-Super†** (trained with CT-Report and CT-Mask pairs) in malignant tumor detection (malignant vs no-tumor). **(d) Performance of R-Super† in tumor detection, malignant tumor detection, and malignancy identification.** Detailed results in Supp. D.1. Error bars represent 95% confidence intervals, calculated with bootstrapping (1,000 bootstrap samples).

2.2 Scale mask datasets

R-Super allows training a tumor segmentation model directly from radiology reports, without tumor masks. In turn, this trained segmentation model can help radiologists to create tumor masks faster: R-Super creates masks, and radiologists (with access to radiology and pathology reports) verify and correct these AI-made masks. R-Super is then fine-tuned on these verified CT-Mask pairs, plus all CT-Report pairs, further improving performance and accelerating annotation. This forms a report-based active learning loop (human-in-the-loop strategy). Traditional active learning has a cold-start problem: it needs the manual creation of many masks, from scratch, before the segmentation model becomes accurate enough to help radiologists. R-Super avoids this by first training on reports alone, then progressively improving as more masks become available. Additionally, we proposed comparing AI-made masks and reports to identify potential AI errors and prioritize these cases for correction. Our report-based active learning strategy is explained in Supp. C.2.

With R-Super-based active learning, 31 radiologists created 1,420 masks for the R-Super Train dataset and 870 masks for UCSF Train (details in Tab. 1). With the support of R-Super, tumor mask creation was 5 times faster for small tumors and 3 times faster for large tumors (Supp. C.2). To the best of our knowledge, our dataset provides the first public CT tumor masks for bladder, gallbladder, spleen, and prostate tumors. For uterus, esophagus, and adrenal tumors, existing public CT datasets contain at most 53 masks per tumor type [19, 20], whereas our training dataset provides from 176 to 288 masks for each of these tumor types (Tab. 1).

Even when CT-Mask pairs are available, CT-Report pairs substantially improve tumor detection. Training R-Super with CT-Report and CT-Mask pairs provides substantially better performance than training with just CT-Mask pairs. We trained R-Super with all CT-Report pairs in UCSF train (116,514), plus different numbers of CT-Mask pairs (0, 175, 350, and 870). We compared it to a standard segmentation model trained with only these CT-Mask pairs (no report), using the same training parameters and AI architecture as R-Super (MedFormer). Figure 3a (detailed results in Fig. D4) shows that R-Super's malignant tumor detection AUC significantly ($p<0.05$) and substantially surpassed the segmentation model both when few masks and when many masks were available in training. Furthermore, the R-Super model trained with all available reports and masks, named R-Super†, substantially surpassed the segmentation model trained with all masks (no reports) on all test sets (UCSF, Stanford, Medipol, and Basel), in tumor detection, malignant tumor detection, and small tumor detection (Supp. D.7). For instance, in malignant tumor detection, R-Super† surpassed the segmentation model by +11% to +13% average sensitivity across the 4 test sets, with both models evaluated at 90% specificity (Tab. D32 to D35). The performance of R-Super† is shown in Fig. 3c and 3d, and detailed in Supp. D.7. In malignant tumor detection, R-Super† achieved from 86% to 91% sensitivity and 88 to 92% specificity (Fig. 3c).

CT-Report pairs substantially improve tumor localization. R-Super localizes tumors by segmenting them. Tumor localization allows radiologists to visualize and verify the AI's predictions. We evaluated tumor localization by comparing R-Super†'s tumor segmentation outputs to 602 tumor masks drawn by our collaborating radiologists on the external Stanford test set (Tab. 1). R-Super†, trained on both masks and reports (UCSF Train), significantly outperformed a segmentation model trained on masks alone in tumor localization: at 175, 350, and 870 training masks, R-Super† achieved +14%, +16%, and +12% higher localization-adjusted sensitivity (Figure 3 and Fig. D5) and +8%, +15%, and +14% higher DSC (Fig. D6) in Stanford Test. Localization-adjusted sensitivity measures the percentage of scans where the AI tumor segmentation output overlaps with radiologist-made tumor masks (evaluated at 80% specificity in Fig. 3). DSC quantifies the amount of overlap between the AI tumor segmentation output and the radiologist-made tumor masks. Overall, results show that including reports in the training dataset substantially improves tumor localization and segmentation.

Training on many reports surpasses training on fewer masks. In tumor detection, R-Super trained on 116,514 reports and no mask surpassed the segmentation model trained on 175, 350, and 870 masks and no report (Fig. 3, external Stanford Test set). In tumor localization, R-Super (no mask) surpassed the segmentation model trained on 175 and 350 masks, in both localization adjusted sensitivity (Fig. 3) and DSC (Fig. D6). Therefore, large-scale report supervision can surpass smaller-scale mask supervision. Importantly, dataset size is not enough for a report-based model to surpass a mask-based model, training strategy is also critical. For instance, a classification model trained on 116,514 reports performed below a segmentation model trained on 175 masks in tumor detection, and much below R-Super (Fig. 3). This classification model was trained on the same dataset as R-Super, using classification labels extracted from all reports (tumor vs no-tumor labels, per-organ), and its architecture was similar to R-Super (MedFormer encoder). Its lower performance shows that a training framework that effectively learns from reports is essential for scaling tumor detection performance. With R-Super, training on reports alone can substitute for training on hundreds of radiologists-drawn masks. When both are available, combining CT-Report and CT-Mask pairs is optimal (R-Super†).

R-Super performance scales with reports. We trained R-Super models with 1,000, 10,000, and 116,514 CT-Report pairs from UCSF Train. More CT-Report pairs in training improved the tumor detection and localization performance, as shown in Fig. 3 (Stanford Test set). This performance improvement was seen both when

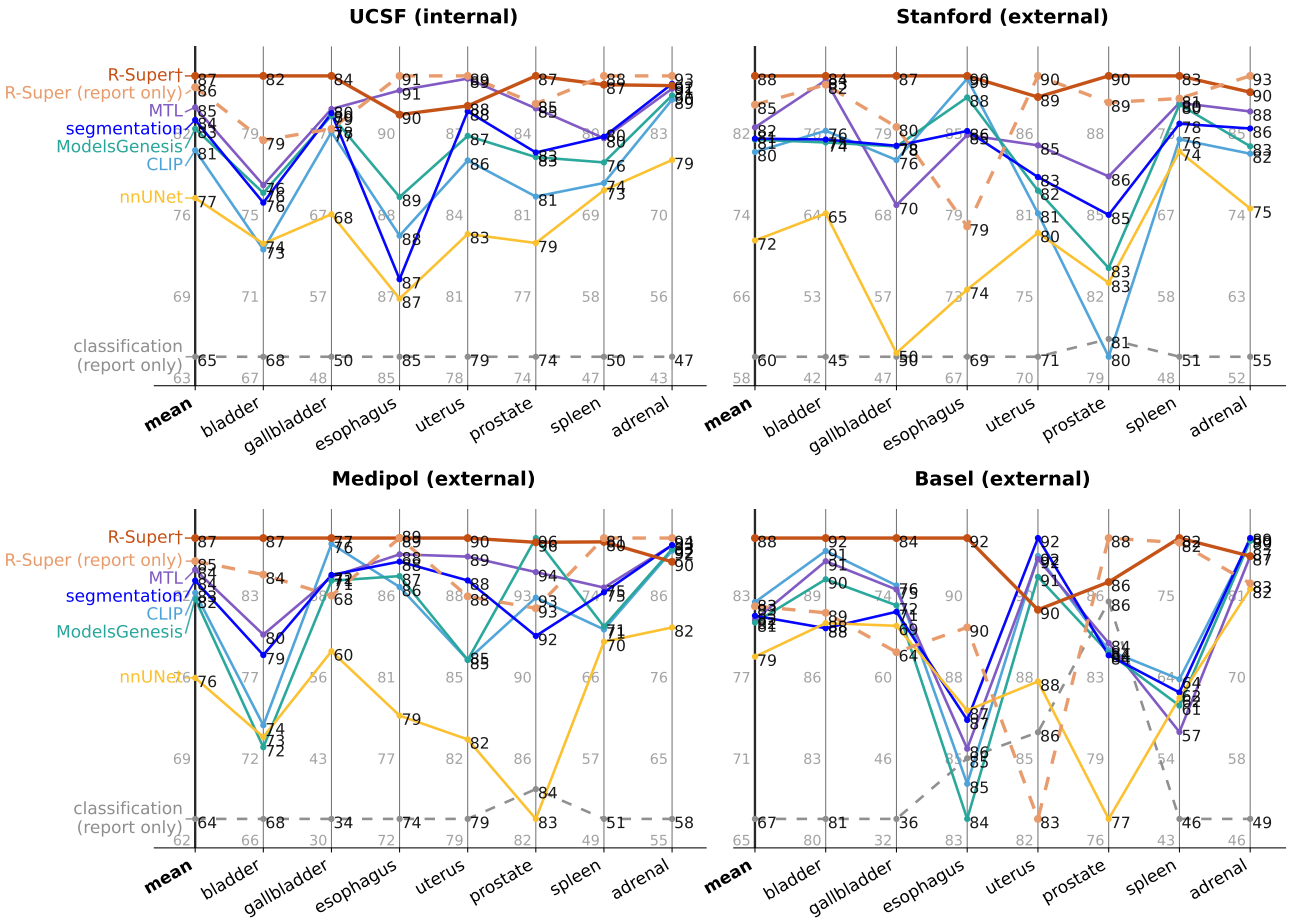

Fig. 4: **By effectively learning from reports, R-Super surpasses alternative methods trained on the same dataset**. The plot compares the tumor detection AUC of R-Super to six alternative training methods. R-Super surpassed all alternative methods on average, on all datasets. All methods were trained on the same data as R-Super, using all available data for its supervision type: we include methods that learn form reports only (R-Super and classification), reports & masks (R-Super†, CLIP, and MTL), masks only (segmentation and nnU-Net), and unlabeled images (Models Genesis pre-trains with CT only, fine-tunes with CT-Mask pairs). We trained on UCSF Train for validation in Stanford, and on the R-Super Train dataset for validation on the other 3 test sets. Dataset details are in Tab. 1 and A1.

training R-Super with CT-Report pairs alone and when training with CT-Report plus CT-Mask pairs. The performance improvement was not saturated in the Fig. 3, even at 116,514 training CT-Report pairs. This result suggests that R-Super can continue improving as larger multi-hospital CT-Report datasets become available.

2.3 Surpass alternative training methods

We compared R-Super to 6 alternative training methods trained on the datasets we created. CLIP: contrastive pretraining on CT-Report pairs followed by CT-Mask fine-tuning for segmentation [30]; multi-task learning (MTL): joint segmentation and classification training from CT-Mask and CT-Report pairs [35]; Models Genesis: self-supervised CT-only pretraining followed by CT-Mask fine-tuning for segmentation [36]; nnU-Net: standard segmentation training from CT-Mask pairs [37, 38]; segmentation: standard segmentation with MedFormer architecture [28]; and classification: classification from report-derived labels (tumor vs no-tumor, per-organ).

Many of these alternative training methods can also learn from CT-Report pairs. However, they do not use reports to directly supervise segmentation, unlike R-Super. All alternative training methods used the same training datasets and segmentation architecture as R-Super (MedFormer [28]), and similar training procedures, except for nnU-Net, which used its automatically defined training parameters and the ResEncL architecture [38]. Regarding architecture modifications, the classifier was based on MedFormer's encoder, MTL attached a classification head to the MedFormer's output features, and all other training methods trained MedFormer without modifications. Each method was trained with all available data for its supervision type: CLIP and MTL used CT-Report and CT-Mask pairs; Models Genesis used all CT scans and CT-Mask pairs; report-based classification used CT-Report pairs; and standard segmentation and nnU-Net used CT-Mask pairs.

R-Super† surpassed all alternative training methods in tumor detection, achieving significantly higher ($p<0.05$) average AUC in all test sets (Fig. 4, Supp. D.7). Against the second-best training method (MTL), R-Super† improved DSC by +14% (Supp. D.2) on Stanford, and improved tumor detection sensitivity by +12%

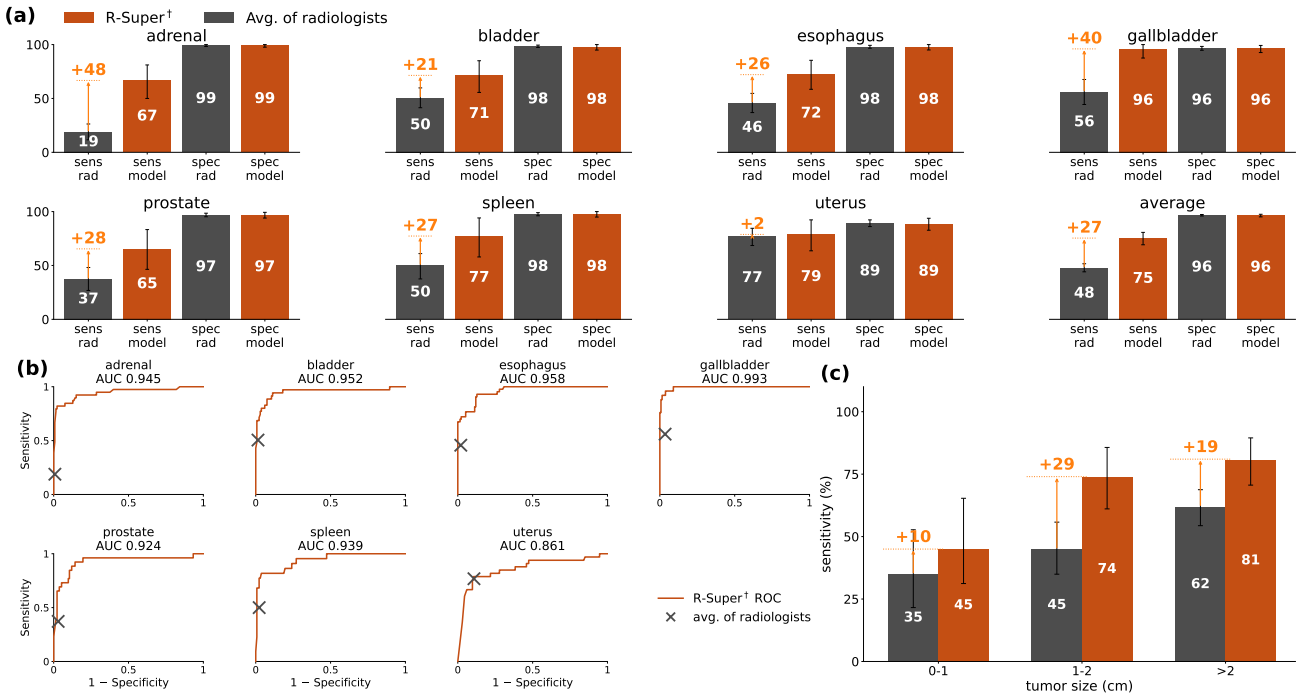

Fig. 5: **R-Super† detected 56% more malignant tumors than the average radiologist (75% vs. 48% average sensitivity) at the same false-positive rate (96% specificity).** Therefore, R-Super† has potential to assist radiologists detect tumors that would have been missed in clinical practice. **(a) R-Super† surpassed the average sensitivity of 6 radiologists in malignant tumor detection for 6 tumor types, and matched the radiologists for uterus tumors.** R-Super† was trained with reports and masks, and evaluated at approximately the same specificity as the radiologist average, on the same reader study dataset (Tab. 1). **(b) ROC curves for R-Super are above the radiologist average for 6 tumor types in malignant tumor detection. (c) R-Super† surpasses radiologists in detecting small malignant tumors.** The bar plot shows sensitivity for R-Super† and the radiologists, for different ranges of tumor diameters. R-Super† surpassed the radiologists for detecting small tumors, especially in the range of 1 to 2 cm. Again, R-Super† was evaluated at approximately the same specificity as the radiologists, the specificity values shown in (a). Error bars represent 95% confidence intervals, calculated with bootstrapping (1,000 bootstrap samples).

on UCSF (internal, Tab. D28), +5% on Stanford (external, Tab. D29), +6% on Medipol (external, Tab. D30), and +10% on Basel (external, Tab. D31), at matched 90% specificity. In malignant tumor detection, R-Super also surpassed all alternative training methods in all datasets (Supp. D.7). By using reports to directly supervise segmentation, R-Super detects tumors better than training methods that use reports to supervise other tasks (e.g., classification or CLIP).

2.4 Surpass radiologists and detect small tumors

To show that AI has potential to assist radiologists, we first need to demonstrate that it can detect tumors they miss. To this end, Supp. B shows that R-Super† substantially surpassed radiologist performance published in the literature for detecting at least 5 tumor types (adrenal, bladder, esophagus, gallbladder, and prostate) and on average. However, comparisons to the literature are limited, since we cannot evaluate R-Super† on the same datasets used to evaluate radiologists in the literature (private datasets).

Therefore, we conducted a reader study, where we evaluated R-Super† and 6 radiologists on the same dataset. It contains 637 CT scans from 513 patients, 50% from UCSF (USA), 30% from Stanford (USA), and 20% from Medipol (Turkey); 132 scans have no tumor, and 505 have tumors, of which 47% include only small tumors (≤ 2 cm in diameter, Tab. A1). This large proportion of small tumors makes this dataset challenging, by focusing on early cancer detection. Additionally, this reader study includes 50% of external data, from Stanford and Medipol. When externally evaluating on Stanford and Medipol data, we used R-Super† trained on UCSF Train. When internally evaluating on UCSF data, we used R-Super† trained on R-Super Train. Dataset details in Tab. 1. All cancer patients had mentions of known (already diagnosed) active cancer in their radiology reports or clinical history, or had positive pathology reports. Because the reader study dataset is large, each radiologist independently read half of the dataset (randomly selected). As a result, each CT scan was read by three radiologists, and radiologist performance is averaged across these three reads. The results for each individual read are in Supp. D.8. The readers are six radiologists with 2 to 8 years of experience (Tab. D40).

For each scan, radiologists answered two questions: (Q1) which organs contained a tumor, generically defined as any focal lesion, and (Q2) which organs contained a malignant tumor. The reader study instructions are in

Supp. H. We used Q1 to evaluate tumor detection and malignant tumor detection. For tumor detection (benign + malignant vs no-tumor), radiologists were correct if they detected any tumor in a scan with a benign or malignant tumor. For malignant tumor detection (malignant vs no-tumor), radiologists were correct if they detected a tumor in a scan with a malignant tumor, regardless of whether they called it malignant; scans with only benign tumors were excluded. We used Q2 to evaluate malignancy identification (malignant vs benign + no-tumor): radiologists were correct if they called a malignant tumor malignant, and did not call any tumor malignant in scans with no tumor or only benign tumors.

In Q1, radiologists marked detected tumors as probable or certain. For the probable level, radiologists were instructed to include any tumor they would mention in a radiology report, simulating which tumors would be detected in clinical practice. In Q2, they also marked malignancy as probable or certain. In this section, we focus on the probable level, considering a tumor (or malignancy) detected if the radiologist marked it as probable. Our primary goal is to test whether AI can reduce completely missed tumors. Once detected, the existence and malignancy of uncertain tumors can be further confirmed through follow-up imaging or other examinations. Detailed results at the certain level are in Supp. D.8.1. At the certain level, radiologists achieved very high average specificity in malignant tumor detection (99%, 95% CI: 98–99%) but low sensitivity (41%, 37–45%). At the same 99% specificity, R-Super† significantly ($p<0.05$) increased sensitivity by +25% on average, achieving sensitivity of 66% (CI: 59-72%, Tab. D50).

R-Super† detected 56% more malignant tumors while maintaining the same specificity as the radiologists. Figure 5 compares the average performance of all radiologists to R-Super†, trained with reports and masks, for malignant tumor detection (malignant vs no-tumor). Evaluated at approximately the same specificity as the radiologists (96%, probable level), R-Super† achieved 75% sensitivity (69-81%) versus 48% (44-52%) for the radiologists, a +27% absolute improvement or +56% relative improvement in sensitivity ($p<0.05$). R-Super's sensitivity was higher for all seven tumor types and significantly higher ($p<0.05$) for six (all except uterus, Tab. D42). Figure 5 also shows tumor detection sensitivity by tumor size. The largest gains of R-Super over the radiologist performance occurred for small tumors between 1 and 2 cm in diameter (+29% sensitivity), suggesting that R-Super has potential to assist radiologists in early cancer detection.

R-Super† also improved tumor detection and malignancy identification. In tumor detection, R-Super's sensitivity surpassed the radiologists ($p<0.05$) for 6 tumor types (all except uterus) at matched specificity (Tab. D41). On average, R-Super† surpassed radiologists by +22% absolute sensitivity at matched specificity ($p<0.05$), reaching 61% average sensitivity (57-65%) at 96% specificity (95-98%). In malignancy identification (malignant vs benign + no-tumor), R-Super† surpassed the average radiologists sensitivity for 5 tumor types ($p<0.05$ for 3), at matched specificity (Tab. D43). The radiologists did not significantly surpass R-Super† for any tumor type (Tab. D43). On average, R-Super† reached 53% sensitivity (46-59%) at 95% specificity in malignancy identification, significantly surpassing radiologists by +14% sensitivity at matched specificity. Tables detailing all reader study results are in Supp. D.8.

Even without tumor masks, R-Super often matched or exceeded radiologist performance. R-Super trained with reports alone surpassed the radiologist sensitivity for 5 tumor types in tumor detection (significant for 2), at matched specificity (Tab. D44). It significantly surpassed the radiologists on average, reaching 46% sensitivity (43-50%) vs 39% radiologist sensitivity (37-41%), at 96% specificity (95-98%). In malignant tumor detection, report-only R-Super surpassed the radiologists' sensitivity in 4 tumor types (significant for 2, at matched specificity) and performed similarly on average (+5% sensitivity, but $p > 0.05$; Tab. D45). In malignancy identification, report-only R-Super surpassed radiologists' sensitivity for 4 tumor types ($p<0.05$ for 2), and performed similarly on average (+3% sensitivity at matched specificity, but $p>0.05$; Tab. D46).

Overall, this reader study demonstrated that R-Super and R-Super† can match or surpass radiologists in a dataset focusing on small tumors. Therefore, the R-Super framework has the potential to assist radiologists detect early tumors that would have been missed.

3 Discussion

This study shows that routine radiology and pathology reports can serve as direct supervision for tumor segmentation. We introduce R-Super, a training framework that converts report-derived tumor descriptions into localized training signals, enabling AI models to learn tumor segmentation without tumor masks. The core advance is to use reports not merely as image-level labels or auxiliary text, but as descriptions of what the segmentation should contain: the number, size, volume, attenuation, rough anatomical location, and malignancy status of tumors. Because reports are already produced at scale in routine care, whereas tumor masks are scarce, expensive, and absent for many tumor types, R-Super changes the scaling logic of tumor segmentation. It enables AI models to learn segmentation from hundreds of thousands of CT-Report pairs, reduces dependence on masks, and provides a scalable path toward more comprehensive, interpretable, and accurate multi-tumor detection.

R-Super is the first public AI model shown to match or surpass radiologists in detecting 7 tumor types on CT: adrenal glands, bladder, esophagus, gallbladder, prostate, spleen, and uterus tumors (Fig. 2). Trained on 127,496 3D CT scans (42 million slices), R-Super substantially outperformed 9 prior public AI models, including public

segmentation models (FLARE and ULS) and leading VLMs from Google, Microsoft, Stanford, and Alibaba (Fig. 2). Usually, reports are used to train VLMs, which can analyze medical images and write AI-made reports. However, these VLM-written reports missed most tumors in our test datasets (Fig. 2), as in previous studies [10, 39]. VLMs performed better in zero-shot classification than in report generation, but R-Super still surpassed them by large margins (Fig. 2). This could be attributed to a design difference. Current VLM architectures and training methods, such as CLIP [40], were originally designed for natural images and captions, which are often vague image descriptions. Medical reports contain rich tumor descriptions, including tumor measurements and approximate locations. The R-Super loss functions convert these descriptions into localized supervision for segmentation. This localized supervision helps R-Super detect tumors that can occupy less than 0.0001% of a CT.

R-Super's performance scales with the number of training reports. We did not observe scaling saturation despite training with over 100,000 reports, representing 27 years of multi-tumor CT scans collected at UCSF and affiliated institutions (Fig. 3a). However, this scaling benefit heavily depends on the AI framework. A classifier trained on the same reports as R-Super, using tumor presence/absence labels extracted from the reports, performed substantially worse than R-Super (Fig. 4). These results show that large-scale report datasets require training frameworks that effectively exploit the detailed information contained in reports. Additionally, this large-scale training was computationally efficient: we trained R-Super in 10 days on one NVIDIA H100 GPU. R-Super also works with any segmentation architecture and adds no inference-time cost.

R-Super can exploit reports and masks together. Thirty-one radiologists created 870 tumor masks for our UCSF Train dataset, with the support of R-Super (Sec. C.2). Training with these CT-Mask pairs plus the CT-Report pairs substantially outperforms training with the CT-Mask pairs alone (Sec. 2.2). R-Super also surpasses previous methods that learned from reports and masks. CLIP [30] pre-trains AI models to pair CT scans and reports. Afterwards, the AI model is fine-tuned for tumor segmentation with CT-Mask pairs. Multi-task learning (MTL, [35]) trains AI models for classification using report-based labels, while masks train the model for segmentation. However, both MTL and CLIP use reports to train the AI for auxiliary tasks, not to directly supervise tumor segmentation. In contrast, R-Super's loss functions directly supervise segmentation, by comparing segmented tumors to tumor descriptions in reports and penalizing discrepancies. This direct supervision allows R-Super† to significantly surpass the alternative methods trained on the same dataset (Fig. 4).

R-Super could help radiologists detect cancer, according to our reader study with 6 radiologists. R-Super† surpassed radiologists in malignant tumor detection for 6 tumor types and matched them for uterus tumors (Sec. 2.4). On average, R-Super† detected 56% more malignant tumors than radiologists at matched specificity, on a dataset focused on small tumors (Fig. 5). These results suggest that R-Super could help radiologists detect small tumors that would otherwise be missed. The tumor types we studied are notably difficult to detect on CT: their signs are subtle, often imperceptible to the human eye. AI can perceive these signs, as previous studies demonstrated for individual tumor types [7], and our study demonstrates for multi-tumor detection.

R-Super performs in a clinically relevant range relative to established cancer-detection exams, though direct comparison is limited and should be read as context rather than equivalence. For spleen, gallbladder, and adrenal tumors, detection already depends largely on incidental findings in imaging, making AI-assisted detection on routine CT a natural objective that could potentially extend to other modalities such as MRI. For the remaining tumor types, specialized cancer-detection exams exist: cystoscopy and cytology for bladder, transvaginal ultrasound (TVS) for uterus, prostate-specific antigen (PSA) and digital rectal examination (DRE) for prostate, and endoscopy for esophagus. For prostate cancer, PSA achieves 21%/94% sensitivity/specificity (4.1 ng/mL cutpoint) [41], while DRE achieves 51%/59% [42], both below the performance of R-Super in our CT test sets (65%/96% sensitivity/specificity in prostatic malignancy identification, Tab. D43). For bladder cancer, urine cytology achieves 10 to 51% sensitivity and 83 to 88% specificity [43], also considerably below R-Super (57% sensitivity and 98% specificity, Tab. D43). However, this comparison is inherently limited: the reference cancer-detection exams were evaluated in populations different from our test datasets, and sensitivity and specificity are not directly transportable. Cystoscopy, endoscopy, and TVS surpass both R-Super and radiologists (cystoscopy: 97% sensitivity, 96% specificity [44]; endoscopy: 94% sensitivity, 92% specificity [45]; TVS: 81% sensitivity, 86% specificity [46]). However, we do not suggest that CT should replace dedicated cancer-detection exams. CT is performed at very large scale for many indications, registering unsuspected tumors that often remain undetected [5]. In contrast, high-accuracy cancer-detection exams are often used in smaller scale and prompted by symptoms or other risk factors. We therefore envision R-Super as an assistive tool that opportunistically flags suspicious findings for clinician review, who, when appropriate, can prompt targeted follow-up with established cancer-detection exams or interval imaging.

Whether this opportunistic use produces net clinical benefit will depend on model calibration. The R-Super performance in Fig. 3d correspond to false-positive rates of 8 to 12%, which would likely create a substantial downstream workup burden (e.g., additional imaging, endoscopy, and short-interval follow-up) if applied to the large, low-cancer-prevalence population undergoing routine CT. Adrenal, prostate, and splenic lesions are recognized sources of overdiagnosis, and detecting additional small or indolent lesions may not improve outcomes and could lead to overtreatment. Therefore, clinical deployment would need low false-positive rates. In our reader study, R-Super operated at 99% specificity while still significantly surpassing radiologists by +25% sensitivity in

malignant tumor detection (Tab. D50), suggesting that false-positive burden may be reduced substantially while preserving clinically meaningful sensitivity gains. However, future prospective evaluation must measure false-positive burden, downstream resource use, target population (general vs high-risk patients), and the proportion of flagged tumors that are clinically actionable. Additionally, R-Super is best positioned as a flagging system for clinician adjudication rather than autonomous diagnosis.

This study has limitations. First, tumor segmentation DSC was lower than is commonly reported for tumor types where CT is the primary detection modality (Supp. D.2). This likely reflects the intrinsic difficulty of outlining our studied tumor types on CT: many are small, subtle, poorly marginated, or poorly visible in CT. The same difficulty appears between humans: when different radiologists annotate these tumors, the DSC between their annotations is also relatively low (Supp. D.2). Nonetheless, R-Super† surpassed the standard mask-based segmentation model by +14% DSC and accurately detected and localized tumors (Fig. 3b), even when contours were imperfect. This aligns with the primary goal of this study: accurate and interpretable tumor detection through segmentation, rather than radiotherapy-grade tumor contouring. Second, this study was retrospective. Although R-Super was evaluated internally and externally across multiple institutions in three continents, prospective multi-center studies are required before clinical implementation. Thus, the present study establishes technical feasibility, external generalization, and reader-study performance, but not yet prospective clinical utility. Third, R-Super uses an LLM to extract tumor information from reports during training. The LLM is highly accurate (up to 98% accuracy, Supp. D.4) but not perfect; yet R-Super achieves high tumor detection and segmentation performance and surpasses alternative training methods despite occasional LLM errors. Fourth, in our reader study, radiologists did not have access to the patients' clinical history because this history often documented known cancers. This setup also simulates incidental detection, where patients often do not present with tumor-related symptoms or history. Previous reader studies on incidental detection used a similar no-history setup [7]. Finally, R-Super's loss functions use organ segmentation masks during training, currently obtained from a public organ segmentation model [10]. Such public models cover dozens of organs [10, 29, 47], but not every organ, and their masks are also not perfect. R-Super has mechanisms to resist moderate organ mask errors (Sec. 4), and it substantially outperforms alternative training methods despite these errors (Fig. 4).

In conclusion, the R-Super results show that routine medical reports can become a scalable open-source foundation for multi-cancer detection and segmentation. Segmentation provides an accurate, interpretable, and verifiable framework to cancer detection, but it has been constrained by the need for voxel-wise tumor masks, which are expensive to create and unavailable for many tumor types. R-Super addresses this constraint by converting medical reports into localized training signals, allowing segmentation models to learn from CT-Report datasets orders of magnitude larger than CT-Mask datasets. Large-scale report supervision surpassed smaller-scale mask supervision for tumor detection and further improved performance when reports and masks were combined (Fig. 3a). Therefore, reports can substitute or supplement masks. By releasing the R-Super framework, trained models, CT-Report pairs, and new tumor masks, this study provides a public foundation for broader multi-cancer detection and segmentation. More broadly, it suggests that cancer imaging may undergo the same transition that has transformed other areas of AI: from small, manually annotated datasets toward large-scale learning from text that is readily available.

4 Methods

R-Super is a training framework designed to learn tumor segmentation from radiology and pathology reports. Unlike tumor masks, reports do not provide voxel-wise tumor locations. However, reports are available in large-scale and provide detailed tumor descriptions: tumor count, diameters, rough locations (organ/slice), attenuation, and malignancy. Turning reports into effective and direct supervision for training tumor segmentation models required a new training framework (Fig. 1): first, R-Super extracts the tumor descriptions from each report using an LLM (Sec. 4.2); then, R-Super trains a tumor segmentation model with four new loss functions, which enforce that segmented tumors match these tumor descriptions extracted from reports:

Volume Loss (L_{vol}, Sec. 4.3): minimizes the dissimilarity between the tumor volume estimated from the segmentation output, and the tumor volume estimated from the radiology report.

Ball Loss (L_{ball}, Sec. 4.4): enforces segmented tumors to match the radiology report descriptions of tumor count, sizes, and rough locations. Given these report descriptions and the output of the segmentation model, the Ball Loss estimates the most probable per-voxel location of each reported tumor, and then maximizes tumor probabilities at those per-voxel locations. The paradigm of estimating a missing variable (here, per-voxel tumor locations) from an observable variable (the report) and a model, and using this estimate to update the model, is theoretically grounded on the Expectation-Maximization framework [48] (Sec. 4.7).

Attenuation Loss (L_{att}, Sec. 4.5): enforces segmented tumors to match the tumor attenuation in radiology reports (i.e., if tumors appear dark or bright).

Malignancy Loss (L_{mal}, Sec. 4.6): enforces segmented tumors to match the malignancy (cancer or not) described in pathology reports, when available.

The four losses combine into the R-Super Loss, L_{rsuper}, show in Eq. 1 (using weights $\lambda_{\text{vol}}, \lambda_{\text{ball}}, \lambda_{\text{att}}, \lambda_{\text{mal}}$, set to 0.1, 0.1, 0.01, and 0.1 in all experiments). In the equation, $\mathbf{x}$ is the 3D CT scan, $\mathbf{r}$ the report, and $f(\mathbf{x};\theta_s)$ the output of the tumor segmentation model. Together, these losses exploit detailed report information but tolerate reports with missing details. Reports are used only during training, never at inference. If CT-Mask pairs are also available for training, we use standard segmentation losses (Dice and cross-entropy, Eq. 2) on CT-Mask pairs $(\mathbf{x}, \mathbf{y} \sim \mathcal{D}_{\text{mask}})$, and the R-Super loss on CT-Report pairs $(\mathbf{x}, \mathbf{r} \sim \mathcal{D}_{\text{report}})$.

For tumor detection, the segmentation output, $f(\mathbf{x};\theta_s)$, is processed by a classifier, $c(f(\mathbf{x};\theta_s);\theta_c)$, which outputs per-organ probabilities of tumor, benign tumor, and malignant tumor. The classifier is trained with cross-entropy loss (L_{cls}) on classification labels ($\mathbf{l}$) extracted from reports (Supp. C.1). Performing tumor detection based on segmentation improves accuracy and explainability (Fig. 3). The R-Super full training objective is shown in Eq. 3. Data pre-processing and training details are presented in Supp. E.

$$\begin{aligned} L_{\text{rsuper}}\big(f(\mathbf{x};\theta_s),\mathbf{r}\big) = \lambda_{\text{vol}} L_{vol}\big(f(\mathbf{x};\theta_s),\mathbf{r}\big) + \lambda_{\text{ball}} L_{ball}\big(f(\mathbf{x};\theta_s),\mathbf{r}\big) \\ + \lambda_{\text{att}} L_{att}\big(f(\mathbf{x};\theta_s),\mathbf{r}\big) + \lambda_{\text{mal}} L_{mal}\big(f(\mathbf{x};\theta_s),\mathbf{r}\big) \end{aligned} \tag{1}$$

$$L_{\text{seg}}\big(f(\mathbf{x};\theta_s),\mathbf{y}\big) = \text{Dice}\big(f(\mathbf{x};\theta_s),\mathbf{y}\big) + \text{CE}\big(f(\mathbf{x};\theta_s),\mathbf{y}\big) \tag{2}$$

$$\min_{\theta_s,\theta_c} \{\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_{\text{mask}}}\big[L_{\text{seg}}\big(f(\mathbf{x};\theta_s),\mathbf{y}\big)\big] + \mathbb{E}_{(\mathbf{x},\mathbf{r})\sim\mathcal{D}_{\text{report}}}\big[L_{\text{rsuper}}\big(f(\mathbf{x};\theta_s),\mathbf{r}\big)\big] + \mathbb{E}_{(\mathbf{x},\mathbf{l})\sim\mathcal{D}}[L_{\text{cls}}(c(f(\mathbf{x};\theta_s);\theta_c),\mathbf{l})]\} \tag{3}$$

The R-Super loss functions leverage organ masks to locate the organs with reported tumors. To create these organ masks, we used an organ segmentation model, an nnU-Net [37] trained in AbdomenAtlas [10] (made public). Organ masks are pre-saved before training the tumor segmentation model. Although public segmentation models can segment few tumor types, they can segment dozens of organs accurately [29, 47]. To compensate for occasional errors in organ masks and account for tumors that grow beyond organs, we dilate these masks by 2 cm. R-Super trains any segmentation architecture, here, we used MedFormer [28], a U-Net-based CNN-transformer hybrid that won previous benchmarks [29].

4.1 A large-scale CT-Report Dataset

Table 1: **Our multi-tumor CT datasets.** We created two training datasets and four test datasets to train and evaluate R-Super. More than 22,000 CT-Report pairs will be publicly released. We directly collected data from hospital PACS systems in three countries: UCSF (USA), Medipol (Turkey), and Basel (Switzerland). Stanford data (25,494 scans) were obtained from the public Merlin dataset [22]. For internal validation on UCSF and external validation on Medipol and Basel, models were trained on R-Super train, which combines UCSF and Stanford data. For external validation on Stanford, models were trained on UCSF train, which excludes Stanford data. The reader study dataset includes cases from UCSF (internal validation), Medipol (external validation), and Stanford (external validation).

	R-Super train	**UCSF train**	**UCSF test**	**Stanford test**	**Medipol test**	**Basel test**	**reader study**
CT scans	127,496	116,514	2,301	1,976	1,327	2,935	637
patients	58,258	43,164	939	1,689	577	2,229	513
controls	13,584	10,614	113	114	178	54	132
source							
country	USA	USA	USA	USA	Turkey	Switzerland	USA & Turkey
hospitals	UCSF & affiliates (80.3%) Stanford (19.7%)	UCSF & affiliates	UCSF & affiliates	Stanford	Medipol	Basel	UCSF & affiliates (50.1%) Stanford (30.1%) Medipol (19.8%)
age and sex							
mean age (SD)	57.7 (18.0)	59.4 (17.0)	62.2 (15.7)	61.9 (16.7)	53.8 (15.4)	68.5 (13.2)	61.5 (15.7)
female, no. (%)	31,557 (54.2)	22,856 (53.0)	445 (47.4)	996 (59.0)	259 (44.9)	1,017 (45.6)	260 (50.8)
male, no. (%)	26,701 (45.8)	20,308 (47.0)	494 (52.6)	693 (41.0)	318 (55.1)	1,212 (54.4)	253 (49.2)
spacing							
in-plane, mm	0.82 (0.47)	0.84 (0.50)	0.83 (0.46)	0.76 (0.09)	0.82 (0.09)	0.79 (0.25)	0.80 (0.28)
slice thickness, mm	2.00 (1.57)	2.23 (1.69)	2.25 (1.90)	1.14 (0.18)	1.43 (0.48)	1.05 (0.30)	1.75 (1.23)
contrast phase							
venous, no. (%)	91,956 (72.1)	69,925 (60.0)	1,494 (64.9)	1,941 (98.2)	1,128 (85.0)	920 (31.3)	516 (81.0)
arterial, no. (%)	8,691 (6.8)	8,827 (7.6)	187 (8.1)	19 (1.0)	136 (10.2)	192 (6.5)	30 (4.7)
non-contrast, no. (%)	12,833 (10.1)	13,553 (11.6)	0 (0.0)	2 (0.1)	0 (0.0)	39 (1.3)	0 (0.0)
unknown, no. (%)	14,016 (11.0)	24,209 (20.8)	620 (26.9)	14 (0.7)	63 (4.7)	1,784 (60.8)	91 (14.3)
tumors							
bladder	4,065	4,050	345	209	146	295	86
gallbladder	3,780	3,562	265	200	174	76	65
prostate	2,527	2,458	180	105	35	170	72
uterus	6,516	5,792	302	330	132	269	104
esophagus	2,720	3,141	388	40	189	95	78
spleen	14,050	13,223	474	494	357	691	95
adrenal gland	11,955	11,477	667	801	357	1,417	127
masks							
total	1,420	870	–	602	–	–	–
bladder	180	88	–	95	–	–	–
gallbladder	141	97	–	76	–	–	–
prostate	156	90	–	68	–	–	–
uterus	288	206	–	91	–	–	–
esophagus	241	191	–	51	–	–	–
spleen	238	142	–	101	–	–	–
adrenal gland	176	56	–	120	–	–	–

We assembled two training and five test datasets, allowing external validation in 3 countries (USA, Turkey, and Switzerland). Datasets are summarized in Tables 1 and A1. Together, the datasets include 134,476 different 3D CT scans and radiology reports, over 5 times the largest public CT datasets [21, 22]. Data was sourced from the PACS system of multiple hospitals: UCSF and affiliated institutions in California, USA; Medipol University Hospital, Turkey; and University Hospital Basel, Switzerland. Additionally, 25,494 CT scans were sourced from the public Merlin dataset (Stanford hospital) [22].

The datasets cover tumor types with no or minimal public CT masks: esophagus, bladder, gallbladder, spleen, uterus, prostate, and adrenal tumors. Only 53 public masks exist for adrenal tumors [19] (adrenocortical carcinoma only), 17 for esophagus tumors, and 39 for uterus tumors [20]. No public CT masks exist for the remaining tumor types. For each tumor type, our training dataset has from 2,527 to 14,050 CT-Report pairs. All data were de-identified prior to this study, and include chest, abdomen, and pelvic CT. Each patient may have multiple time points and contrast phases (Tab. 1).

Each dataset contains no-tumor cases, benign tumors, and malignant tumors (counts in Tab. A1). Here, benign tumors are generically defined as non-malignant focal lesions, encompassing cysts and solid lesions, identified by explicit benign characterization in radiology reports and absence of malignancy (in the benign tumor's organ) across all available ICD-10 codes, radiology reports, and pathology reports. Malignant tumors include primary cancers and metastases (Tab. A1). In UCSF Test, malignancy was confirmed using pathology reports. In the remaining test datasets, where pathology reports were unavailable, malignancy was confirmed by explicit malignancy characterization in radiology reports. In the reader study dataset, malignancy was determined from pathology reports (UCSF data, 50%), or by explicit mentions of already diagnosed active cancer in radiology reports or clinical history (e.g., "Findings consistent with known cervical carcinoma", or ongoing cancer treatment). Test control patients (no-tumor patients) had no mention of a current or prior tumor in any radiology/pathology report or diagnostic code, for all 7 organs we focus on. Control patients may nonetheless present with other abnormalities such as infection or aneurysm, which is important for realistic evaluation: R-Super must distinguish tumors from other pathologies, not only from healthy tissue. The training dataset contains pathology reports for part of the data (Tab. A1).

We will publicly release over 22,000 CT-Report pairs, and the first public tumor segmentation masks for multiple tumor types (Tab. 1). These masks were created by 31 radiologists assisted by R-Super, in an active learning loop (Sec. C.2). More details about each dataset, data acquisition, and train-test splits are in Supp. A.

4.2 LLM for Extracting Report Information

We use an LLM to extract tumor characteristics from reports. From radiology reports, it extracts tumor count, rough locations (organ/tumor slice), attenuation, and diameters. From pathology reports, it extracts whether tumors are malignant or benign. For training cases without pathology reports, the LLM looks for high-confidence confirmations of malignancy in the radiology report instead.

LLMs can extract information from reports created by different radiologists, in different styles, since LLMs can understand language semantics and context [10]. To keep the LLM applicable to different report styles and avoid overfitting to the reports in our training dataset, we use a zero-shot LLM, Qwen 3 30B (A3B Thinking FP8) [49]. To reduce computational cost, we run the LLM only once per report and store its answer. Our LLM prompt (Supp. G) was iteratively developed by radiologists and computer scientists through testing and error analysis. The prompt provides the LLM with medical knowledge and clear guidelines for interpreting reports. It also asks the LLM to justify its answers and fill a template with tumor characteristics. The template is automatically converted into a table, used as ground-truth for the Volume, Ball, Attenuation, and Malignancy Losses.

The LLM achieved 98.3% (95% CI: 97.6-98.9) accuracy and 95.8% F1-Score (94-97.4%) in determining whether tumors exist in each organ (i.e., determining tumor existence and rough location), 94.1% (91.2-96.7%) accuracy in determining malignancy, 95.6% (93.1-98.0) in determining size, and 92.9% (86.2-98.6%) in determining attenuation. The LLM evaluation is in Supp. D.4. This LLM performance surpasses methods used to extract information from reports and create ground-truth labels in popular medical datasets (e.g., 90% F1-Score in ChestX-ray8 [50]). Nevertheless, the LLM is not perfect, and occasional LLM errors exist in the R-Super training process. Despite this, R-Super substantially outperformed other training methods and public AI models (Fig. 4,2), indicating robustness to moderate LLM errors.

4.3 Volume Loss

Radiology reports do not inform the exact tumor contour, needed for usual segmentation loss functions. However, from reports we can estimate tumor volumes. The Volume Loss teaches the segmentation model to segment tumors that match tumor volumes estimated from the report. Specifically, it enforces two constraints: segmented tumors must be in the *rough locations* (organs) described in the report; and, for each rough location (o), the *combined volume* of all segmented tumors ($V_{s,o}$) must match the combined volume of all tumors in the report ($V_{r,o}$). We estimate the segmented tumor volume ($V_{s,o}$) by summing per-voxel tumor probabilities (output of the segmentation model) within the rough location o (organ). We estimate the reported tumor volume ($V_{r,o}$)

from the tumor diameters described in the radiology report. The Volume Loss is designed to be easy to optimize and converge, as it only constrains overall tumor volumes, not tumor diameters or counts. We apply it to an intermediate layer of the segmentation model (e.g., the second decoder layer in MedFormer) to produce a rough segmentation. The remaining losses, applied to the last layer, then teach the segmentation model to refine this rough segmentation to match multiple tumor characteristics in the report. The Volume Loss is shown in Fig. 1.

Overall equation ($L_{\mathbf{vol},o}$). Equation 4 summarizes the Volume Loss for an organ o. When tumors are reported in the organ, the term $L_{\text{forg},o}(V_{s,o}, V_{r,o})$ minimizes the dissimilarity between the segmented tumor volume $V_{s,o}$ and the reported tumor volume $V_{r,o}$ in the organ o. Meanwhile, the term $L_{\text{bkg},o}(\mathbf{T}^o)$ penalizes tumor segmentations outside the organ o (e.g., spleen tumors segmented outside the spleen). $\mathbf{T}^o$ is the segmentation model output channel for tumors in the organ o. For an organ (o) where the report mentions no tumor, we simply penalize the entire tumor segmentation output ($\mathbf{T}^o$) using per-voxel cross-entropy with a zero target (Eq. 4). The overall volume loss (L_{vol}, Eq. 5) is the sum of the per-organ volume loss $L_{\text{vol},o}$ for all organs of interest ($o \in \mathcal{O}$). Next, we explain the calculation of each term in Eq. 4.

$$L_{\text{vol},o} = \begin{cases} L_{\text{forg},o}(V_{s,o}, V_{r,o}) + L_{\text{bkg},o}(\mathbf{T}^o), & \text{if tumors are reported in } o \\ \text{CE}(\mathbf{T}^o, \mathbf{0}), & \text{otherwise} \end{cases} \tag{4}$$

$$L_{\text{vol}} = \sum_{o \in \mathcal{O}} L_{\text{vol},o} \tag{5}$$

Estimating tumor volumes from reports ($V_{r,o}$). Most reports do not include tumor volumes. They usually include tumor diameters, from which we estimate volumes. For each tumor, 1, 2 or 3 perpendicular diameters are provided. From the diameters, we estimate tumor volumes. With 1 tumor diameter (d_1), we estimate tumor volume as a the volume of a ball ($d_1^3\pi/6$); with 3 diameters (d_1, d_2, d_3), we use the volume of an ellipsoid $d_1 d_2 d_3 \pi/6$; with 2 diameters (d_1 and d_2), we estimate the third as $d_3 = (d_1 + d_2)/2$, and use the ellipsoid volume. For each organ o, we estimate the volume of all tumors in the organ and sum them, estimating the reported tumor volume $V_{r,o}$.

Estimating tumor volumes from segmentation ($V_{s,o}$) and background penalization ($L_{\mathbf{bkg},o}$). We estimate $V_{s,o}$ from the segmentation model's output $\mathbf{T}^o = [t^o_{h,w,l}]$, the per-voxel tumor probabilities for tumors in organ o. Inside $\mathbf{T}^o$, we first locate the organ o, using an organ segmentation mask ($\mathbf{O} = [o_{h,w,l}]$; organ masks are AI-made, pre-saved, and dilated to compensate for errors and tumors growing beyond organs). Outside the organ, the background penalization term ($L_{\text{bkg},o}$) minimizes the tumor probabilities estimated by the AI, using cross-entropy with target of 0 (Eq. 6, where $\odot$ stands for element-wise multiplication). To estimate $V_{s,o}$, we sum all per-voxel tumor probabilities inside the organ, and multiply them by the voxel volume v (Eq. 7). If tumor probabilities were binary (0/1), Eq. 7 would be equivalent to counting the tumor voxels inside the organ. In reality, tumor probabilities $t^o_{h,w,l}$ are continuous, so $V_{s,o}$ is a soft estimation of tumor volume.

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

$$V_{s,o} = v \sum_{h,w,l}^{H,W,L} t^o_{h,w,l} o_{h,w,l} \tag{7}$$

Volume dissimilarity function and tolerating report errors ($L_{\mathbf{forg},o}$). The Volume Loss minimizes the dissimilarity between the segmented tumor volume ($V_{s,o}$) and the reported tumor volume ($V_{r,o}$), by minimizing the dissimilarity function $L_{\text{forg},o}(V_{s,o}, V_{r,o})$ in Eq. 8. In the equation, E is a constant (set to 500 mm^3) that provides numerical stability when $V_{r,o}$ is small. This function, plotted in Fig. 1, surpassed standard loss functions (e.g., L1 and L2 losses) in preliminary tests, and it has 3 design principles. First, its measures a dissimilarity between $V_{s,o}$ and $V_{r,o}$: its minimum is when $V_{s,o}$ matches $V_{r,o}$, and it increases with the difference between $V_{s,o}$ and $V_{r,o}$. Second, it has a strong (but finite) gradient when $V_{s,o}$ is zero but $V_{r,o}$ is not. Therefore, the Volume Loss strongly penalizes the segmentation model for missing tumors mentioned in the report (false negatives). To further prevent false negatives, gradients are softer when $V_{s,o}$ is larger than $V_{r,o}$. Third, $L_{\text{forg},o}(V_{s,o}, V_{r,o})$ has a tolerance margin (created by the max operation in Eq. 8): if the difference between the segmented tumor volume and the reported tumor volume is small (we use a tolerance, τ, of 10%), $L_{\text{forg},o}(V_{s,o}, V_{r,o})$ and its gradient become zero, not penalizing the AI. We use this tolerance because $V_{r,o}$ is not a perfect measurement of the true tumor volume. Our estimation of $V_{r,o}$ from tumor diameters in the report is an approximation. Furthermore, the diameters in the report are subject to inter-observer variance (two radiologists measuring the same tumor often provide slightly different diameters). The Volume Loss tolerance (τ) compensates for these inaccuracies.

$$L_{\text{forg},o}(V_{s,o}, V_{r,o}) = \max\{L'_{\text{forg},o}(V_{s,o}, V_{r,o}) - L'_{\text{forg},o}((1-\tau)V_{r,o}, V_{r,o}), 0\} \tag{8}$$

$$L'_{\text{forg},o}(V_{s,o}, V_{r,o}) = \frac{|V_{s,o} - V_{r,o}|}{V_{s,o} + V_{r,o} + E} \tag{9}$$

Exploiting tumor slices and reports missing tumor sizes or counts. Reports may also provide the tumor slice, i.e., the vertical coordinate z_i of a tumor i in organ o. When slices are provided, the Volume Loss becomes more precise, teaching the AI to only segment tumors near these slices. To this end, before applying the Volume Loss, we edit the organ mask ($\tilde{\mathbf{O}} \leftarrow \mathbf{O}$) by setting it to zero at vertical coordinates more than one tumor diameter (d_i) away from every reported slice z_i (Eq. 10, where $\mathcal{Z} = \{z_i\}$ is the set of slices for all tumors in organ o, and x, y, z are spatial coordinates). This restriction makes the segmented tumor volume $V_{s,o}$ depend only on regions near the reported slices (Eq. 7), encouraging tumor segmentation there. Meanwhile, the $L_{\text{bkg},o}$ term of the Volume Loss (Eq. 6) penalizes segmentations far from the slices. The Volume Loss can also handle reports missing tumor sizes or tumor counts (e.g., reports that only measure and count the largest tumors). In these cases, we employ a prior-based, high-tolerance reformulation of the Volume Loss, explained in Supp. C.3.

$$\tilde{o}(x,y,z) = \begin{cases} 0, & \text{if } |z - z_i| > d_i \ \ \forall z_i \in \mathcal{Z} \\ o(x,y,z), & \text{otherwise} \end{cases} \tag{10}$$

Cropping requirement. Medical segmentation models are typically trained on CT crops rather than whole scans. The Volume Loss needs the crop to fully contain the organ with tumors. Otherwise, the Volume Loss could encourage the model to segment a tumor that is not inside the crop. We randomly choose on which organ to crop on, giving higher priority (90%) to organs with tumors. If a second organ with tumor is partially inside the crop, its tumor channels receive no loss.

4.4 Ball Loss

Radiology reports inform not only approximate tumor volumes, but also diverse tumor characteristics. The Ball Loss teaches the segmentation model to segment tumors that match the radiology report in tumor count, diameters, and rough locations (organs / slice). Thus, the Ball Loss is more strict than the Volume Loss, and it is applied to the last layer of the segmentation model. In summary, the Ball Loss uses the output of the tumor segmentation model and the tumor descriptions in radiology reports to locate the most probable position for each tumor described in the report. In each position, the Ball Loss optimizes the segmented tumor to match the corresponding tumor size (diameter and volume) from the report. As an example (Fig. 1), consider a report mentions a 30 mm spleen tumor. The Ball Loss first locates the spleen. Inside the spleen, it locates the 30 mm sphere where the tumor probability (predicted by the segmentation model) is highest. Inside this sphere, the Ball Loss maximizes the top-N most probable voxels, where N is the tumor volume estimated from the report (Sec. 4.3), in voxels, minus a tolerance margin (20%). Around these top-N most probable voxels, an uncertainty margin is not optimized, compensating for uncertainty at tumor borders. Outside this margin, tumor probabilities are minimized. Therefore, the Ball Loss enforces the segmentation of a single spleen tumor of any shape that fits inside a sphere of 30 mm, the tumor diameter in the report. The Ball Loss is summarized in Fig. 1.

Overall Equation ($L_{\text{ball},o}$). For organs o where the report mentions no tumor, the Ball Loss minimizes the AI tumor segmentation output ($\mathbf{T^o}$), by minimizing a cross-entropy with target 0 (Eq. 11). For organs with reported tumors, we define the Ball Loss as $L'_{\text{ball},o}$, calculated with three operations: ball convolution, finding of the most likely tumor voxels, and maximization of tumor probabilities in these voxels. These operations are explained below. The final Ball Loss, aggregated over all organs of interest ($o \in \mathcal{O}$), is shown in Eq. 12.

$$L_{\text{ball},o} = \begin{cases} L'_{\text{ball},o}, & \text{if tumors are reported in } o \\ \text{CE}(\mathbf{T}^o, \mathbf{0}), & \text{otherwise} \end{cases} \tag{11}$$

$$L_{\text{ball}} = \sum_{o \in \mathcal{O}} L_{\text{ball},o} \tag{12}$$

Ball Convolution: locating the most likely tumor positions. First, we use pre-saved organ masks $\mathbf{O}$ (AI-made organ masks dilated by 2 cm) to locate the organ o in the tumor segmentation output $\mathbf{T}^o$ (per-voxel tumor probabilities). Inside the organ o, we locate each tumor described in the report. Initially, consider that organ o has a single reported tumor i. To locate it, we use an operation that we called *Ball Convolution*: a standard 3D convolution (stride 1 and zero-padding), but with a non-learnable kernel shaped like a ball ($\mathbf{k}_i$,

defined in Eq. 13). This ball is binary (1 inside, 0 outside) and has the same diameter d_i as the tumor diameter in the report (the largest reported diameter for tumor i). We apply the Ball Convolution to the tumor segmentation output (tumor probabilities, $\mathbf{T}^o$), multiplied by the organ mask ($\mathbf{O}$), as shown in Eq. 14, where $*$ stands for convolution. The Ball Convolution moves the ball (kernel) over the tumor segmentation output. At each position, the ball sums all the tumor probabilities inside it. Therefore, the maximum output of the Ball Convolution locates the most probable center ($\mathbf{c}$) for a tumor with the reported diameter d (Eq. 15, where x, y, z are spatial coordinates). I.e., a ball of diameter d and center $\mathbf{c}$ represents the most probable per-voxel tumor location, the *highest probability ball* ($\mathbf{H}$), defined in Eq. 16. In the equation, μ represents a small dilation (we set $\mu = 20\%$), used to compensate for inaccuracies in the reported diameter d. The Ball Convolution cannot set the highest probability ball center $\mathbf{c}$ outside the organ o, due to the previous multiplication between the tumor segmentation output and the organ mask (Eq. 14).

$$\mathbf{k}_i(u,v,w) = \begin{cases} 1, & \text{if } \sqrt{u^2+v^2+w^2} \leq \dfrac{d}{2} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

$$\mathbf{B} = (\mathbf{O} \odot \mathbf{T}^o) * \mathbf{k}_i \tag{14}$$

$$\mathbf{c} = [c_x, c_y, c_z] = \underset{(x,y,z)}{\operatorname{argmax}} \mathbf{B}(x,y,z) \tag{15}$$

$$\mathbf{H}(x,y,z) = \begin{cases} 1, & \text{if } \|(x,y,z) - \mathbf{c}\| \leq \dfrac{d(1+\mu)}{2} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

Locating the most likely tumor voxels. The Ball Loss teaches the AI to segment a tumor that fits inside the highest probability ball ($\mathbf{H}$), but this tumor does not need to be spherical. To allow for various tumor shapes, we do not maximize all voxels inside the highest probability ball, but only its top-N voxels with the highest tumor probabilities. We define N as the tumor volume estimated from the report (as in Sec. 4.3), reduced by a small tolerance margin (20%), and converted to a voxel count (N). The top-N voxels can assume any shape that fits inside the highest probability ball, and they have a reasonable volume, considering the tumor volume in the report. The top-N voxels represent the most likely tumor voxels given the segmentation output and the tumor description in the radiology report (tumor organ and diameter).

Maximization and minimization of tumor probabilities. The Ball Loss maximizes the tumor probabilities in these top-N most probable tumor voxels. To this end, we define a segmentation mask $\mathbf{M}$, where the top-N voxels are set to 1, and other voxels to 0 (Eq. 17). We use this mask as the target for cross-entropy and Dice losses applied to the segmentation model's output $\mathbf{T}^o$. However, to account for the segmentation model's uncertainty at tumor borders, we introduce an uncertainty margin consisting of the next most probable voxels within the highest probability ball, i.e., the top-$[N+1, \nu N]$ voxels (we set $\nu = 1.2$). This margin is stored in an uncertainty mask $\mathbf{M}'$ (Eq. 18). We skip penalization of these uncertain voxels by masking the segmentation model's output $\mathbf{T}^o$ with $(1 - \mathbf{M}')$ before the CE and Dice losses, preventing gradient propagation and optimization in this uncertain region. The Ball Loss with this uncertainty margin is defined in Eq. 19. The masks $\mathbf{M}$ and $\mathbf{M}'$ are generated continuously in training, improving as the segmentation model improves.

$$\mathbf{M}(x,y,z) = \begin{cases} 1, & \text{if } (x,y,z) \in \operatorname{top}_{[1,N]} \left(\mathbf{T}^o(x,y,z) \odot \mathbf{H}(x,y,z) \right) \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

$$\mathbf{M}'(x,y,z) = \begin{cases} 1, & \text{if } (x,y,z) \in \operatorname{top}_{[N+1,\nu N]} \left(\mathbf{T}^o(x,y,z) \odot \mathbf{H}(x,y,z) \right) \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

$$L'_{\text{ball},o} = \text{CE}((1-\mathbf{M}') \odot \mathbf{T}^o, \mathbf{M}) + \text{Dice}((1-\mathbf{M}') \odot \mathbf{T}^o, \mathbf{M}) \tag{19}$$

Reports with multiple tumors. When reports mention multiple tumors in an organ o, we locate them from largest to smallest. For each tumor i, we repeat the previously explained procedure: a Ball Convolution identifies the tumor's highest-probability ball, the top-N most probable voxels inside this ball are added to the mask $\mathbf{M}_i$, and the top-$[N+1, \nu N]$ voxels are added to the uncertainty mask $\mathbf{M}'_i$. After each step, these selected top-N voxels are removed from the tumor segmentation output ($\mathbf{T}^o$) so they cannot be reused for the next tumor. After we locate all tumors reported in organ o, we compute the union of their individual masks, generating the joint masks $\mathbf{M}$ and $\mathbf{M}'$ (Eq. 20). These joint masks become the target for the cross-entropy and Dice losses in Eq. 19.

$$\mathbf{M} = \bigcup_i \mathbf{M}_i, \qquad \mathbf{M}' = \bigcup_i \mathbf{M}'_i \tag{20}$$

Exploiting tumor slices and reports missing tumor sizes or counts. The Ball Loss leverages the available information reports, becoming more precise when reports provide more tumor details. When reports include tumor slices, the Ball Loss teaches the AI to segment tumors at the reported slices. To this end, we adapt each Ball Convolution to locate a tumor near a reported slice, by temporarily zeroing the organ mask in slices more than 1 tumor diameter away from the reported tumor slice (Eq. 10, but setting $\mathcal{Z}$ to a single slice z_i for each Ball Convolution, $\mathcal{Z} = [z_i]$). This makes the Ball Convolution search region very small, making the Ball Loss very precise. The Ball Loss is also usable for less precise reports that lack the number or diameters of tumors. In these cases, we use a high-tolerance reformulation of the Ball Loss, which teaches the segmentation model to segment at least the number of tumors explicitly reported in the radiology report (see Supp. Note C.4).

4.5 Attenuation Loss

The Attenuation Loss constrains tumor segmentation to be consistent with the tumor attenuation described in radiology reports (Fig. 1). Attenuation encodes the relative brightness of the tumor with respect to the surrounding organ (hypo-, hyper-, or isoattenuating). To calculate the Attenuation Loss, we first estimate brightness statistics. Specifically, the mean and standard deviation of the organ voxels and the tumor voxels in the CT (μ^o, σ^o, $\mu^{t,o}$, and $\sigma^{t,o}$, respectively). These four statistics are given to a small attenuation classifier (MLP with 128 hidden neurons). It classifies whether the tumors are hypoattenuating, hyperattenuating, or mixed/isoattenuating. The label for the attenuation classifier is extracted from the report by the LLM. We train the classifier with a standard cross-entropy classification loss. The gradient of this loss trains the attenuation classifier, but it also back-propagates to the segmentation model, because the statistics $\mu^{t,o}$ and $\sigma^{t,o}$ depend on the tumor voxels, which are defined according to the segmentation model's output $\mathbf{T}^o$ (Eq. 23 to 25). The gradient encourages the segmentation model to help the attenuation classifier: it should segment tumors whose attenuation match the report, so that the attenuation classifier can correctly classify their attenuation. The Attenuation Loss is applied both as deep supervision and at the last layer of the segmentation model.

Formulation. The Attenuation Loss for organ o is shown in Eq. 21, where the attenuation classifier is represented as function $a(\mu^o, \sigma^o, \mu^{t,o}, \sigma^{t,o}; \theta_a)$, with parameters θ_a. In case there is no tumor reported in the organ o or when reports do not inform attenuation, the loss is set to 0; otherwise, it its set as the attenuation classifier cross-entropy loss. The attenuation label, extracted from report, is encoded as an integer $A \in 0, 1, 2$ (representing hypo-, hyper-, and iso-/mixed attenuation, where mixed indicates that different tumors in organ o have different attenuation). Eq. 22 shows the loss aggregated for all organs of interest ($o \in \mathcal{O}$). Organ and tumor brightness statistics ($\mu^o, \sigma^o, \mu^{t,o}, \sigma^{t,o}$) are defined in Eq. 23 and 24. These statistics are calculated from the CT scan $\mathbf{X}$ (normalized per-voxel HU values), the segmentation model's output $\mathbf{T}^o$ (tumor probabilities), and the pre-saved organ segmentation mask $\mathbf{O}$.

$$L_{att,o} = \begin{cases} \mathrm{CE}(a(\mu^o, \sigma^o, \mu^{t,o}, \sigma^{t,o}; \theta_a), A), & \text{if tumors are reported in } o \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

$$L_{att} = \sum_{o \in \mathcal{O}} L_{att,o} \tag{22}$$

$$\mu^o = \frac{\sum_{h,w,l} m^o_{h,w,l}\, x_{h,w,l}}{\sum_{h,w,l} m^o_{h,w,l} + \varepsilon}, \qquad \sigma^o = \sqrt{\frac{\sum_{h,w,l} m^o_{h,w,l}\, (x_{h,w,l} - \mu^o)^2}{\sum_{h,w,l} m^o_{h,w,l} + \varepsilon}} \tag{23}$$

$$\mu^{t,o} = \frac{\sum_{h,w,l} t^o_{h,w,l}\, x_{h,w,l}}{\sum_{h,w,l} t^o_{h,w,l} + \varepsilon}, \qquad \sigma^{t,o} = \sqrt{\frac{\sum_{h,w,l} t^o_{h,w,l}\, (x_{h,w,l} - \mu^{t,o})^2}{\sum_{h,w,l} t^o_{h,w,l} + \varepsilon}} \tag{24}$$

$$m^o_{h,w,l} = (1 - t^o_{h,w,l})\, o_{h,w,l} \tag{25}$$

4.6 Malignancy Loss

The Malignancy Loss (L_{mal}) trains the segmentation model to identify tumors as benign or malignant. From pathology reports, the LLM extracts whether each tumor is malignant (cancer) or benign. Thus, by using the LLM to analyze pathology and radiology reports together, we define the diameter, rough location (organ / slice), and malignancy of each tumor. For CT scans with tumors of inconclusive malignancy (e.g., a benign and a malignant tumor in the same organ with similar sizes or an inconclusive report), we skip the Malignancy Loss. For scans without pathology reports, we extract malignancy from the radiology report instead, if it includes malignancy information (e.g., radiology reports often mention the patient has a known active cancer). The segmentation model includes separate output channels for segmenting general tumors, malignant tumors, and benign tumors (3 channels per organ). The Volume, Ball, and Attenuation Losses train the tumor channels, while the benign and malignant channels are trained exclusively by the Malignancy Loss.

Formulation. The Malignancy Loss is a variation of the Ball Loss (Sec. 4.4), beginning with the same procedure: Ball Convolutions are applied to general tumor segmentation output of the segmentation model ($\mathbf{T}^o$) to identify, for each reported tumor, the most probable tumor voxels given the reported tumor size and rough location; these voxels define the tumor mask $\mathbf{M}_i$, while the surrounding voxels (uncertain tumor borders) define the uncertainty mask $\mathbf{M'}_i$ (Eq. 13 to 18). Afterwards, unlike the Ball Loss, the Malignancy Loss merges the masks for benign and malignant tumor separately, producing the $\mathbf{M}^b$ and $\mathbf{M'}^b$ joint masks for benign tumors, and $\mathbf{M}^m$ and $\mathbf{M'}^m$ for malignant tumors, as defined in Eq. 26. In the equation, $\mathcal{B}^o$ and $\mathcal{M}^o$ denote the sets of benign and malignant tumors in organ o, respectively, as determined from reports. The joint masks for benign and malignant tumors ($\mathbf{M}^b$ and $\mathbf{M}^m$) are used as targets for Dice and CE losses, which supervise the segmentation model's outputs for benign and malignant tumors ($\mathbf{T}^{o,b}$ and $\mathbf{T}^{o,m}$), as shown in Eq. 27 and 28. As in the Ball Loss, uncertainty masks ($\mathbf{M'}^b$ and $\mathbf{M'}^m$) define voxels that are not optimized. If organ o contains no benign / malignant tumor (and no tumor of unknown malignancy), the corresponding benign / malignant segmentation output is minimized with a cross-entropy loss with target zero (Eq. 29 and 30). The final Malignancy Loss over all organs of interest ($o \in \mathcal{O}$) is given in Eq. 32. The benign and malignant tumor masks ($\mathbf{M}^m, \mathbf{M}^b$) that guide the Malignancy Loss are derived from the same general tumor output $\mathbf{T}^o$ (and reports), enforcing consistency in the segmentation model outputs for benign, malignant, and overall tumors.

$$\mathbf{M}^b = \bigcup_{i \in \mathcal{B}^o} \mathbf{M}_i, \quad \mathbf{M'}^b = \bigcup_{i \in \mathcal{B}^o} \mathbf{M'}_i, \quad \mathbf{M}^m = \bigcup_{i \in \mathcal{M}^o} \mathbf{M}_i, \quad \mathbf{M'}^m = \bigcup_{i \in \mathcal{M}^o} \mathbf{M'}_i \tag{26}$$

$$L'^b_{\text{mal},o} = \text{CE}\big((1 - \mathbf{M'}^b) \odot \mathbf{T}^{o,b}, \mathbf{M}^b\big) + \text{Dice}\big((1 - \mathbf{M'}^b) \odot \mathbf{T}^{o,b}, \mathbf{M}^b\big) \tag{27}$$

$$L'^m_{\text{mal},o} = \text{CE}\big((1 - \mathbf{M'}^m) \odot \mathbf{T}^{o,m}, \mathbf{M}^m\big) + \text{Dice}\big((1 - \mathbf{M'}^m) \odot \mathbf{T}^{o,m}, \mathbf{M}^m\big) \tag{28}$$

$$L^b_{\text{mal},o} = \begin{cases} L'^b_{\text{mal},o}, & \text{if benign tumors are reported in } o \\ \text{CE}(\mathbf{T}^{o,b}, \mathbf{0}), & \text{otherwise} \end{cases} \tag{29}$$

$$L^m_{\text{mal},o} = \begin{cases} L'^m_{\text{mal},o}, & \text{if malignant tumors are reported in } o \\ \text{CE}(\mathbf{T}^{o,m}, \mathbf{0}), & \text{otherwise} \end{cases} \tag{30}$$

$$L_{\text{mal},o} = L^b_{\text{mal},o} + L^m_{\text{mal},o} \tag{31}$$

$$L_{\text{mal}} = \sum_{o \in \mathcal{O}} L_{\text{mal},o} \tag{32}$$

4.7 Rationale and Robustness of the Report Supervision Losses

Does the Ball Loss work if the segmentation model cannot find the tumor? The Ball Loss remains effective even when the AI output is weak (i.e., all predicted per-voxel tumor probabilities are low). In this case, the Ball Convolutions still find the most probable tumor voxels, and maximize probabilities in these voxels. If the AI model predicted only zero probabilities, the Ball Convolution would fail. However, this situation is unlikely, because the Volume Loss strongly penalizes the AI when all predicted tumor probabilities are zero but the report mentions tumors. The Volume Loss supports the Ball Loss.

Can the Ball Loss reinforce false positives? In a single training image, if the model segments a false positive inside an organ that contains a reported tumor (i.e., a tumor segmented in the wrong part of the organ), the Ball Loss may temporarily reinforce that false positive. However, this mistake will not be reinforced consistently across the training dataset. When a similar false positive appears in a no-tumor patient, the Ball Loss will penalize it. To minimize the Ball Loss across the whole training dataset, the segmentation model must learn to segment tumors that match multiple tumor characteristics in reports: tumor rough locations, counts, and sizes. The Attenuation and Malignancy Losses add further constraints by requiring segmented tumors to match the reported tumor attenuation and malignancy. False positives do not match all of these characteristics consistently across patients, but true tumors do. In general, loss functions are known to potentially enforce incorrect solutions for individual training samples (e.g., false positive tumors), but minimizing these losses over the full dataset favors solutions that reduce error consistently across samples [51, 52] (e.g., correct tumor segmentations). Consistent with this, our new report-based losses substantially improved tumor detection and localization compared with training on CT-Mask pairs alone (Sec. 2.2). This result indicates that, across the full training dataset, the new losses guide the model toward true tumors rather than false positives.

Relationship to theory on learning from incomplete data. Theoretically, the Ball Loss is inspired by the principles of the Expectation–Maximization (EM) algorithm, a classical statistical framework for learning from incomplete data problems [53]. In such problems, part of the data cannot be directly observed (here, the per-voxel tumor segmentation mask), and another part can be observed (here, report descriptions of tumor counts, diameters, and rough locations). At a high level, EM alternates between estimating the unobserved data from the observed data and the current model (Expectation step), and updating the model so that its outputs better match

the estimated unobserved data. In an adaptation to object segmentation in photographs [54], the Expectation step consisted in estimating the most likely object mask (unobserved data) given a segmentation model's output and image-level binary classification labels (observed data); the Maximization step then optimized the segmentation model to better match this most likely object mask. The Ball Loss follows an EM-like structure: it presents a new algorithm (based on the Ball Convolution) that estimates a most likely tumor mask given the output of a segmentation model and the detailed tumor description in the report; then, it optimizes the segmentation model to better match this most likely tumor mask. As in classical EM, repeating this procedure leads to a progressive improvement of both the model and the estimates of the unobserved data: as the segmentation model improves, the estimated most likely tumor masks improve, resulting in increasingly accurate supervision.

The Volume Loss does not follow an EM-like formulation, as it does not estimate a most likely tumor mask. Instead, it directly calculates the tumor volume from the segmentation model's output and enforces this volume to match the tumor volume estimated from the report. The accuracy of EM-based models is known to depend on initialization [55]. At a high level, applying the Volume Loss to hidden layers of the segmentation model is akin to an EM initialization procedure: it encourages the segmentation model to produce a reasonable coarse tumor mask, which later layers can refine according to the Ball Loss. Our ablation studies show that the Volume Loss is very important for accuracy (Supp. D.5), as initializations are important for EM.

Statistical Analysis. We calculated 95% confidence intervals with bootstrapping with 1,000 samples with replacement. For significance testing we used paired permutation tests, with 10,000 permutations, except for comparisons of AUC. For AUC, we used DeLong's test. All reported P values are two-sided.

Acknowledgments. 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.

Funding. This work was supported by the National Institutes of Health (NIH) under Award Number R01EB037669 and the Lustgarten Foundation for Pancreatic Cancer Research. This research was funded in part by National Science Center, Poland 2025/57/B/ST6/03100. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018792.

Author Contributions. P.R.A.S.B. developed the study's concept and algorithm, acquired and organized data, ran experiments, analyzed results, lead data annotation, and wrote the manuscript. All authors reviewed the manuscript and provided revisions and feedback. X.Z. acquired and organized data, ran experiments, lead data annotation, and analyzed results. W.L. acquired and organized data, lead data annotation, lead reader studies, and provided technical advice. S.P. lead data annotation and reader studies, and provided technical advice. J.W. acquired data for external validation. J.C. provided technical advice. I.E.H. and S.E. acquired data for external validation and lead reader studies. J.P. performed data annotation. Z.Zhu organized data. G.D., W.R., Y.B.T, F.H.M., and M.A. collaborated in reader studies. Y.W. and S.Y. acquired and organized data. Q.C. performed experiments. A.K., B.M., J.B.C, Y.Z., Y.G., A.S.C., C.P.L., S.D. and A.C. provided technical advice. U.B. provided technical advice and lead reader studies. K.W. and Y.Y. acquired and organized data. A.S., K.W., Y.Y., A.L.Y., and Z.Zhou supervised the project and advised on its direction.

Competing Interests. The authors declare no competing interests.

Materials & Correspondence. Correspondence and requests for materials should be addressed to Pedro R. A. S. Bassi, Arkadiusz Sitek, Kang Wang, Yang Yang, Alan L. Yuille, or Zongwei Zhou.

Ethics Statement. All datasets used in this study were collected and analyzed in accordance with institutional guidelines and regulations. The use of retrospective clinical data was approved by the Institutional Review Board at Johns Hopkins University (IRB00403268). External data was obtained with appropriate institutional approvals and data-use agreements. Patient demographic information, including age and sex was derived from EHRs. All data were de-identified before training and analysis, and no protected health information was used in this study. All training and analysis were conducted on servers compliant with Protected Health Information to ensure that patient privacy was fully protected. Access to the data was restricted to individuals with appropriate institutional certifications and training in handling sensitive patient information.

Code Availability. The R-Super custom code is publicly available at GitHub (https://github.com/PedroRASB/SegmentationFromReports). The implementation builds on open-source libraries, including Python (v.3.10) and PyTorch (v.2.10).

Data Availability. This paper is based on large-scale CT datasets, including 127,496 CT scans in R-Super train, 116,514 in UCSF Train, 2,301 in UCSF Test, 1,976 in Stanford Test, 1,327 in Medipol Test, 2,935 in Basel Test, and 637 in the reader study dataset. Each CT scan is paired with its corresponding radiology report. The Stanford data (25,494 CT scans and reports) come from the public Merlin dataset, available at https://stanfordaimi.azurewebsites.net/datasets/60b9c7ff-877b-48ce-96c3-0194c8205c40. The remaining data

was acquired by our team from hospital PACS, converted from DICOM to NifTI, and de-identified by removing all patient-identifiable metadata. We already have authorization to release 22,000 CT scans and the corresponding reports. Authorization to release the remaining data is pending. Besides CT scans and reports, tumor masks created by 31 radiologists will also be made publicly available. The de-identified data will be made available upon publication through HuggingFace under a CC-BY-NC-ND-4.0 license.

References

[1] Mathers, C.D., Boerma, T., Ma Fat, D.: Global and regional causes of death. British medical bulletin **92**(1), 7–32 (2009)

[2] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)

[3] Crosby, D., Bhatia, S., Brindle, K.M., Coussens, L.M., Dive, C., Emberton, M., Esener, S., Fitzgerald, R.C., Gambhir, S.S., Kuhn, P., Rebbeck, T.R., Balasubramanian, S.: Early detection of cancer. Science **375**(6586) (2022) https://doi.org/10.1126/science.aay9040

[4] McCollough, C.H., Bushberg, J.T., Fletcher, J.G., Eckel, L.J.: Answers to common questions about the use and safety of ct scans. In: Mayo Clinic Proceedings, vol. 90, pp. 1380–1392 (2015). Elsevier

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

[6] Xia, Y., Yu, Q., Chu, L., Kawamoto, S., Park, S., Liu, F., Chen, J., Zhu, Z., Li, B., Zhou, Z., Yuille, A.L., Fishman, E.K., Hruban, R.H.: The felix project: Deep networks to detect pancreatic neoplasms. medRxiv (2022)

[7] Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., Tang, W., Jin, G., Jiang, H., Fang, X., Nogues, I., Li, X., Guo, W., Wang, Y., Fang, W., Qiu, M., Hou, Y., Kovarnik, T., Vocka, M., Lu, Y., Chen, Y., Chen, X., Liu, Z., Zhou, J., Xie, C., Zhang, R., Lu, H., Hager, G.D., Yuille, A.L., Lu, L., Shao, C., Shi, Y., Zhang, Q., Liang, T., Zhang, L., Lu, J.: Large-scale pancreatic cancer detection via non-contrast ct and deep learning. Nature Medicine **29**(12), 3033–3043 (2023) https://doi.org/10.1038/s41591-023-02640-w

[8] Hu, C., Xia, Y., Zheng, Z., Cao, M., Zheng, G., Chen, S., Sun, J., Chen, W., Zheng, Q., Pan, S., Zhang, Y., Chen, J., Yu, P., Xu, J., Xu, J., Qiu, Z., Lin, T., Yun, B., Yao, J., Guo, W., Gao, C., Kong, X., Chen, K., Wen, Z., Zhu, G., Qiao, J., Pan, Y., Li, H., Gong, X., Ye, Z., Ao, W., Zhang, L., Yan, X., Tong, Y., Yang, X., Zheng, X., Fan, S., Cao, J., Yan, C., Xie, K., Zhang, S., Wang, Y., Zheng, L., Wu, Y., Ge, Z., Tian, X., Zhang, X., Wang, Y., Zhang, R., Wei, Y., Zhu, W., Zhang, J., Qiu, H., Su, M., Shi, L., Xu, Z., Zhang, L., Cheng, X.: Ai-based large-scale screening of gastric cancer from noncontrast ct imaging. Nature Medicine **31**(9), 3011–3019 (2025) https://doi.org/10.1038/s41591-025-03785-6

[9] Li, W., Bassi, P.R.A.S., Wu, L., Zhou, X., Zhao, Y., Chen, Q., Plotka, S., Lin, T., Zhu, Z., Martin, M., Caskey, J., Jiang, S., Chen, X., Ćwikła, J.B., Sankowski, A., Wu, Y., Decherchi, S., Cavalli, A., Lall, C., Tomasetti, C., Guo, Y., Yu, X., Cai, Y., Qiao, H., Bao, J., Hu, C., Wang, X., Sitek, A., Ding, K., Li, H., Wang, M., Yu, D., Zhang, G., Yang, Y., Wang, K., Yuille, A.L., Zhou, Z.: Early and prediagnostic detection of pancreatic cancer from computed tomography. arXiv preprint arXiv:2601.22134 (2026)

[10] 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.: Radgpt: Constructing 3d image-text tumor datasets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 23720–23730 (2025). https://github.com/MrGiovanni/RadGPT

[11] Liu, J., Zhang, Y., Wang, K., Yavuz, M.C., Chen, X., Yuan, Y., Li, H., Yang, Y., Yuille, A., Tang, Y., Zhou, Z.: Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography. Medical Image Analysis, 103226 (2024)

[12] Hooper, S., Chen, M., Saab, K., Bhatia, K., Langlotz, C., Ré, C.: A case for reframing automated medical image classification as segmentation. Advances in Neural Information Processing Systems **36**, 55415–55441 (2023)

[13] Bassi, P.R., Dertkigil, S.S., Cavalli, A.: Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization. Nature Communications **15**(1), 291 (2024)

[14] Trimpl, M.J., Primakov, S., Lambin, P., Stride, E.P., Vallis, K.A., Gooding, M.J.: Beyond automatic medical image segmentation—the spectrum between fully manual and fully automatic delineation. Physics in Medicine & Biology **67**(12), 12–01 (2022)

[15] Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., Ginneken, B., Bilello, M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M.J., Heckers, S.H., Huisman, H., Jarnagin, W.R., McHugo, M.K., Napel, S., Pernicka, J.S.G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Meakin, J.A., Ourselin, S., Wiesenfarth, M., ArbelÃ¡ez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, I., Maier-Hein, K., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A.L., Maier-Hein, L., Cardoso, M.J.: The medical segmentation decathlon. Nature Communications **13**(1) (2022) https://doi.org/10.1038/s41467-022-30695-9

[16] Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., LohÃ¶fer, F., Holch, J.W., Sommer, W., Hofmann, F., Hostettler, A., Lev-Cohain, N., Drozdzal, M., Amitai, M.M., Vivanti, R., Sosna, J., Ezhov, I., Sekuboyina, A., Navarro, F., Kofler, F., Paetzold, J.C., Shit, S., Hu, X., LipkovÃ¡, J., Rempfler, M., Piraud, M., Kirschke, J., Wiestler, B., Zhang, Z., HÃ¼lsemeyer, C., Beetz, M., Ettlinger, F., Antonelli, M., Bae, W., Bellver, M., Bi, L., Chen, H., Chlebus, G., Dam, E.B., Dou, Q., Fu, C.-W., Georgescu, B., GirÃ³-i-Nieto, X., Gruen, F., Han, X., Heng, P.-A., Hesser, J., Moltz, J.H., Igel, C., Isensee, F., JÃ¤ger, P., Jia, F., Kaluva, K.C., Khened, M., Kim, I., Kim, J.-H., Kim, S., Kohl, S., Konopczynski, T., Kori, A., Krishnamurthi, G., Li, F., Li, H., Li, J., Li, X., Lowengrub, J., Ma, J., Maier-Hein, K., Maninis, K.-K., Meine, H., Merhof, D., Pai, A., Perslev, M., Petersen, J., Pont-Tuset, J., Qi, J., Qi, X., Rippel, O., Roth, K., Sarasua, I., Schenk, A., Shen, Z., Torres, J., Wachinger, C., Wang, C., Weninger, L., Wu, J., Xu, D., Yang, X., Yu, S.C.-H., Yuan, Y., Yue, M., Zhang, L., Cardoso, J., Bakas, S., Braren, R., Heinemann, V., Pal, C., Tang, A., Kadoury, S., Soler, L., Ginneken, B., Greenspan, H., Joskowicz, L., Menze, B.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023) https://doi.org/10.1016/j.media.2022.102680

[17] Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y., Zhang, Y., Wang, Y., Hou, F., Yang, J., Xiong, G., Tian, J., Zhong, C., Ma, J., Rickman, J., Dean, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Kaluzniak, H., Raza, S., Rosenberg, J., Moore, K., Walczak, E., Rengel, Z., Edgerton, Z., Vasdev, R., Peterson, M., McSweeney, S., Peterson, S., Kalapara, A., Sathianathen, N., Papanikolopoulos, N., Weight, C.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis **67**, 101821 (2021) https://doi.org/10.1016/j.media.2020.101821

[18] 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.: Pants: The pancreatic tumor segmentation dataset. In: Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2025). https://github.com/MrGiovanni/PanTS

[19] Moawad, A.W., Ahmed, A.A., ElMohr, M., Eltaher, M., Habra, M.A., Fisher, S., Perrier, N., Zhang, M., Fuentes, D., Elsayes, K.: Voxel-level segmentation of pathologically-proven adrenocortical carcinoma with ki-67 expression (adrenal-acc-ki67-seg)[data set]. The Cancer Imaging Archive **8** (2023)

[20] 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.: Scaling tumor segmentation: Best lessons from real and synthetic data. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 24001–24013 (2025). https://github.com/BodyMaps/AbdomenAtlas2.0

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

[22] Blankemeier, L., Kumar, A., Cohen, J.P., Liu, J., Liu, L., Van Veen, D., Gardezi, S.J.S., Yu, H., Paschali, M., Chen, Z., Delbrouck, J.-B., Reis, E., Holland, R., Truyts, C., Bluethgen, C., Wu, Y., Lian, L., Jensen, M.E.K., Ostmeier, S., Varma, M., Valanarasu, J.M.J., Fang, Z., Huo, Z., Nabulsi, Z., Ardila, D., Weng, W.-H., Amaro Junior, E., Ahuja, N., Fries, J., Shah, N.H., Zaharchuk, G., Willis, M., Yala, A., Johnston, A., Boutin,

R.D., Wentland, A., Langlotz, C.P., Hom, J., Gatidis, S., Chaudhari, A.S.: Merlin: a computed tomography vision-language foundation model and dataset. Nature (2026) https://doi.org/10.1038/s41586-026-10181-8

[23] Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., Chen, J., Mahvar, F., Yatziv, L., Chen, T., Sterling, B., Baby, S.A., Baby, S.M., Lai, J., Schmidgall, S., Yang, L., Chen, K., Bjornsson, P., Reddy, S., Brush, R., Philbrick, K., Asiedu, M., Mezerreg, I., Hu, H., Yang, H., Tiwari, R., Jansen, S., Singh, P., Liu, Y., Azizi, S., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Riviere, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.-b., Ramos, S., Yvinec, E., Casbon, M., Buchatskaya, E., Alayrac, J.-B., Lepikhin, D., Feinberg, V., Borgeaud, S., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L., Joulin, A., Bachem, O., Matias, Y., Chou, K., Hassidim, A., Goel, K., Farabet, C., Barral, J., Warkentin, T., Shlens, J., Fleet, D., Cotruta, V., Sanseviero, O., Martins, G., Kirk, P., Rao, A., Shetty, S., Steiner, D.F., Kirmizibayrak, C., Pilgrim, R., Golden, D., Yang, L.: MedGemma Technical Report (2026). https://arxiv.org/abs/2507.05201

[24] Team, L., Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., Sun, Y., Shen, J., Wang, C., Tan, J., Zhao, D., Xu, T., Zhang, H., Rong, Y.: Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning (2025). https://arxiv.org/abs/2506.07044

[25] Codella, N.C.F., Jin, Y., Jain, S., Gu, Y., Lee, H.H., Abacha, A.B., Santamaria-Pang, A., Guyman, W., Sangani, N., Zhang, S., Poon, H., Hyland, S., Bannur, S., Alvarez-Valle, J., Li, X., Garrett, J., McMillan, A., Rajguru, G., Maddi, M., Vijayrania, N., Bhimai, R., Mecklenburg, N., Jain, R., Holstein, D., Gaur, N., Aski, V., Hwang, J.-N., Lin, T., Tarapov, I., Lungren, M., Wei, M.: MedImageInsight: An Open-Source Embedding Model for General Domain Medical Imaging (2024). https://arxiv.org/abs/2410.06542

[26] Zhuang, J., Wu, L., Wang, Q., Fei, P., Vardhanabhuti, V., Luo, L., Chen, H.: Mim: Mask in mask self-supervised pre-training for 3d medical image analysis. IEEE Transactions on Medical Imaging (2025)

[27] Zhuang, J., Luo, L., Wang, Q., Wu, M., Luo, L., Chen, H.: Advancing volumetric medical image segmentation via global-local masked autoencoders. IEEE Transactions on Medical Imaging (2025)

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

[29] 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.: Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track **37**, 15184–15201 (2024)

[30] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision (2021). https://arxiv.org/abs/2103.00020

[31] Grauw, M., Scholten, E.T., Smit, E.J., Rutten, M.J., Prokop, M., Ginneken, B., Hering, A.: The uls23 challenge: A baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography. Medical Image Analysis **102**, 103525 (2025)

[32] Ma, J., Zhang, Y., Gu, S., Ge, C., Wang, E., Zhou, Q., Huang, Z., Lyu, P., He, J., Wang, B.: Automatic Organ and Pan-cancer Segmentation in Abdomen CT: the FLARE 2023 Challenge (2024). https://arxiv.org/abs/2408.12534

[33] Ma, J., Wang, B.: Fast, Low-Resource, Accurate Robust Organ and Pan-cancer Segmentation: MICCAI Challenge, FLARE 2024. Springer (2025)

[34] Ma, J., Gu, S., Wang, B.: Fast, low-resource, accurate, robust, and effectual Medical Image Analysis (FLARE) 2025. Zenodo (2025). https://doi.org/10.5281/zenodo.15094799 . https://doi.org/10.5281/zenodo.15094799

[35] Chen, E.Z., Dong, X., Li, X., Jiang, H., Rong, R., Wu, J.: Lesion attributes segmentation for melanoma

detection with multi-task u-net. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 485–488 (2019). IEEE

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

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

[38] Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. arXiv preprint arXiv:2404.09556 (2024)

[39] Chen, Y., Xiao, W., Bassi, P.R., Zhou, X., Er, S., Hamamci, I.E., Zhou, Z., Yuille, A.: Are vision language models ready for clinical diagnosis? a 3d medical benchmark for tumor-centric visual question answering. arXiv preprint arXiv:2505.18915 (2025)

[40] Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

[41] Thompson, I.M., Ankerst, D.P., Chi, C., Lucia, M.S., Goodman, P.J., Crowley, J.J., Parnes, H.L., Coltman, C.A.: Operating characteristics of prostate-specific antigen in men with an initial psa level of 3.0 ng/ml or lower. Jama **294**(1), 66–70 (2005)

[42] Naji, L., Randhawa, H., Sohani, Z., Dennis, B., Lautenbach, D., Kavanagh, O., Bawor, M., Banfield, L., Profetto, J.: Digital rectal examination for prostate cancer screening in primary care: a systematic review and meta-analysis. The Annals of Family Medicine **16**(2), 149–154 (2018)

[43] Yafi, F.A., Brimo, F., Auger, M., Aprikian, A., Tanguay, S., Kassouf, W.: Is the performance of urinary cytology as high as reported historically? a contemporary analysis in the detection and surveillance of bladder cancer. In: Urologic Oncology: Seminars and Original Investigations, vol. 32, pp. 27–1 (2014). Elsevier

[44] Dahmcke, C.M., Steven, K.E., Larsen, L.K., Poulsen, A.L., Abdul-Al, A., Dahl, C., Guldberg, P.: A prospective blinded evaluation of urine-dna testing for detection of urothelial bladder carcinoma in patients with gross hematuria. European Urology **70**(6), 916–919 (2016)

[45] Wong, M.C., Deng, Y., Huang, J., Bai, Y., Wang, H.H., Yuan, J., Zhang, L., Yip, H.C., Chiu, P.W.Y.: Performance of screening tests for esophageal squamous cell carcinoma: a systematic review and meta-analysis. Gastrointestinal Endoscopy **96**(2), 197–207 (2022)

[46] Jacobs, I., Gentry-Maharaj, A., Burnell, M., Manchanda, R., Singh, N., Sharma, A., Ryan, A., Seif, M.W., Amso, N.N., Turner, G., Brunell, C., Fletcher, G., Rangar, R., Ford, K., Godfrey, K., Lopes, A., Oram, D., Herod, J., Williamson, K., Scott, I., Jenkins, H., Mould, T., Woolas, R., Murdoch, J., Dobbs, S., Leeson, S., Cruickshank, D., Skates, S.J., Fallowfield, L., Parmar, M., Campbell, S., Menon, U.: Sensitivity of transvaginal ultrasound screening for endometrial cancer in postmenopausal women: a case-control study within the ukctocs cohort. The Lancet Oncology **12**(1), 38–48 (2011) https://doi.org/10.1016/s1470-2045(10)70268-0

[47] Wasserthal, J., Breit, H.-C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023) https://doi.org/10.1148/ryai.230024

[48] Moon, T.K.: The expectation-maximization algorithm. IEEE Signal processing magazine **13**(6), 47–60 (1996)

[49] 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.: Qwen3 Technical Report (2025). https://arxiv.org/abs/2505.09388

[50] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray

database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106 (2017)

[51] Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics, 400–407 (1951)

[52] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436 (2015)

[53] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22 (1977)

[54] Papandreou, G., Chen, L.-C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1742–1750 (2015)

[55] McLachlan, G.J., Krishnan, T.: The EM algorithm and extensions. John Wiley & Sons (2008)

[56] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison (2019). https://arxiv.org/abs/1901.07031

[57] Malik, R.F., Berry, R., Lau, B.D., Busireddy, K.R., Patel, P., Patel, S.H., Fishman, E.K., Bivalacqua, T.J., Johnson, P.T., Sedaghat, F.: Systematic evaluation of imaging features of early bladder cancer using computed tomography performed before pathologic diagnosis. Tomography **9**(5), 1734–1744 (2023)

[58] Sui, H., Ma, R., Liu, L., Gao, Y., Zhang, W., Mo, Z.: Detection of incidental esophageal cancers on chest ct by deep learning. Frontiers in Oncology **11**, 700210 (2021)

[59] Frezza, E., Mezghebe, H.: Gallbladder carcinoma: a 28 year experience. International surgery **82**(3), 295–300 (1997)

[60] Korevaar, S., Tennakoon, R., Page, M., Brotchie, P., Thangarajah, J., Florescu, C., Sutherland, T., Kam, N.M., Bab-Hadiashar, A.: Incidental detection of prostate cancer with computed tomography scans. Scientific Reports **11**(1) (2021) https://doi.org/10.1038/s41598-021-86972-y

[61] Valizadeh, P., Jannatdoust, P., Tahamtan, M., Ghorani, H., Dorcheh, S.S., Farnoud, K., Salahshour, F.: Diagnostic performance of different imaging modalities for splenic malignancies: A comparative meta-analysis. European Journal of Radiology Open **12**, 100566 (2024)

[62] Franconeri, A., Fang, J., Brook, A., Brook, O.R.: Asymptomatic endometrial thickening of 8 mm or greater on postcontrast computed tomography in postmenopausal women is a predictor of endometrial cancer. Journal of computer assisted tomography **43**(1), 136–142 (2019)

[63] Allard, P., Yankaskas, B.C., Fletcher, R.H., Alden Parkery, L., Halvorsen Jr, R.A.: Sensitivity and specificity of computed tomography for the detection of adrenal metastatic lesions among 91 autopsied lung cancer patients. Cancer **66**(3), 457–462 (1990)

[64] Littlefair, S., Mello-Thoms, C., Reed, W., Pietryzk, M., Lewis, S., McEntee, M., Brennan, P.: Increasing prevalence expectation in thoracic radiology leads to overcall. Academic radiology **23**(3), 284–289 (2016)

[65] Chen, J., Gandomkar, Z., Reed, W.M.: Investigating the impact of cognitive biases in radiologists' image interpretation: a scoping review. European journal of radiology **166**, 111013 (2023)

[66] Qu, C., Zhang, T., Qiao, H., Liu, J., Tang, Y., Yuille, A., Zhou, Z.: 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, vol. 21 (2023). https://github.com/MrGiovanni/AbdomenAtlas

[67] 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.: Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. Medical Image Analysis, 103285 (2024)

Supplementary Information: Contents

A Supplementary Dataset Information **28**

B Comparison of our Reader Study to the Literature **30**
B.1 Details of Selected Studies 31

C Supplementary Methods **31**
C.1 Tumor Detection and Malignancy Identification 31
C.2 R-Super-guided Mask Creation 32
C.3 Volume Loss for Reports without Tumor Size or Count 33
C.4 Ball Loss for Reports without Tumor Size or Count 33
C.5 Details on the Ball Loss 34

D Supplementary Results **34**
D.1 R-Super Results: Summary Tables 34
D.2 Segmentation DSC and Inter-Reader Agreement 40
D.3 Scaling Performance 42
D.4 LLM Evaluation 47
D.5 Ablation Studies 48
D.6 Comparison to Public AI Models 51
D.7 Comparison to Alternative Training Methods Trained on Our Dataset 64
D.8 Comparison to Radiologists 76

E Training Details **88**

F VLM Evaluation **88**

G LLM Prompts **89**

H Instructions for Reader Study **95**

Appendix A Supplementary Dataset Information

This section presents details about each of our training and testing datasets:

(1) UCSF Train (new). LLMs searched over 410,000 radiology and pathology reports from the University of California San Francisco (UCSF) Picture Archiving and Communication System (PACS), spanning 1997–2024 and covering the UCSF hospital and multiple affiliated institutions across California, USA. The LLMs identified control patients and patients with tumors. We included all tumor cases discovered by the LLMs, plus a random selection of controls (total of 116,514 CT-Report pairs). For efficiency, a small LLM (Llama 3.1 8B AWQ) performed initial report screening, followed by confirmation of selected reports by a large LLM (Qwen 3 30B). Both models used radiologist-designed prompts. This LLM and prompt combination was previously validated by radiologists and shown to achieve 96% accuracy in extracting tumor information from reports [10]. This accuracy is comparable to established labelers, such as in CheXpert [56] and ChestX-ray14 [50].

(2) UCSF Test (new). We randomly selected 2,301 CT-Report pairs of UCSF Whole for internal validation. Control patients were randomly selected with age and sex distributions matched to the tumor patients.

(3) R-Super Train (new). We excluded patients from the UCSF test dataset and merged the remaining UCSF CT-Report pairs with all 25,494 CT-Report pairs from Merlin, the largest public abdominal CT dataset (Stanford Hospital, Palo Alto, California, 2012 to 2018) [22]. The merged dataset, R-Super Train, was used to train the R-Super model that was evaluated in UCSF Test (internal), Medipol (external), and Basel (external). We trained a second R-Super model on the whole UCSF dataset (UCSF train, without Stanford data) and externally evaluated it in the Stanford test set and the reader study dataset.

(4) Stanford Test (existing, [22]). The public Merlin dataset, from Stanford, was collected from the Stanford Hospital, from 2012 to 2018. Exams were selected via the Stanford Medicine Research Data Repository (STARR), using Current Procedural Terminology (CPT) codes 72192, 72193, 72194, 74150, 74160, 74170, 74176, 74177, and 74178. For external validation, we selected 1,976 CT-Report pairs from the Merlin official test set, including all cases with malignant tumors in the 7 organs we studied, plus 172 randomly selected controls and 965 benign tumor cases.

(5) Medipol Test (new). We used the same LLMs used in UCSF to analyze reports from the Medipol University Hospital in Istanbul, Turkey, from 2014 to 2025. We included all tumor cases discovered by the LLMs, plus a random selection of controls (total of 1,327 CT-Report pairs). The entire dataset was used for external validation.

(6) Basel Test (new). Diagnostic codes (ICD-10) were used to search for tumor patients across the PACS of the University Hospital Basel, in Basel, Switzerland, from 2010 to 2024. In total, 2,935 CT-Report pairs were selected. The entire dataset was used for external validation.

(7) Reader Study (new). Our reader study dataset includes 637 CT scans, 50% from UCSF, 30% from Stanford, and 20% from Medipol. Reading this number of CT scans is a large, but still possible radiologist workload. We focused this dataset on small tumors: 47% of the scans include only tumors smaller than 2 cm in diameter.

Table A1: **Tumors per dataset.** Here, *rad* and *path* indicate confirmation of tumor type (malignant, metastasis, primary) from radiology reports and pathology reports, respectively. A tumor may be confirmed by both. Benign tumors are confirmed from radiology reports. Radiology-based confirmation includes only radiology reports that indicate high confidence in the tumor type (e.g., known cancer history). *Handling unknown malignancy:* test set tumors of unknown malignancy are included when evaluating for tumor detection (malignant + benign vs no-tumor), and excluded when evaluating for malignant tumor detection (malignant vs no-tumor) or malignancy identification (malignant vs benign + no-tumor).

	R-Super train	UCSF train	UCSF test	Stanford test	Medipol test	Basel test	reader study
bladder tumors							
total	4,065	4,050	345	209	146	295	86
small	808	838	107	41	54	37	49
malignant (rad)	758	735	–	55	39	79	10
small malignant (rad)	152	154	–	7	9	5	5
metastatic (rad)	418	441	–	9	5	6	2
benign	760	684	70	31	35	6	22
unknown malignancy	2,015	1,878	86	123	72	210	29
malignant (path)	532	753	189	–	–	–	25
small malignant (path)	96	154	55	–	–	–	22
primary malignant (path)	377	580	174	–	–	–	25
gallbladder tumors							
total	3,780	3,562	265	200	174	76	65
small	761	706	65	41	75	7	24
malignant (rad)	471	437	–	41	3	17	1
small malignant (rad)	84	89	–	2	0	0	0
metastatic (rad)	192	203	–	6	1	11	3
benign	2,218	2,030	103	77	81	11	16
unknown malignancy	1,036	950	84	82	90	48	24
malignant (path)	55	145	78	–	–	–	24
small malignant (path)	1	10	8	–	–	–	5
primary malignant (path)	49	139	78	–	–	–	24
prostate tumors							
total	2,527	2,458	180	105	35	170	72
small	197	213	31	4	4	14	11
malignant (rad)	409	384	–	21	2	28	3
small malignant (rad)	34	34	–	0	0	0	0
metastatic (rad)	105	120	–	0	0	2	3
benign	718	617	60	38	0	2	6
unknown malignancy	919	843	31	46	33	140	40
malignant (path)	481	614	89	–	–	–	23
small malignant (path)	25	43	15	–	–	–	6
primary malignant (path)	470	603	89	–	–	–	23
uterus tumors							
total	6,516	5,792	302	330	132	269	104
small	827	654	57	43	12	12	38
malignant (rad)	571	539	–	45	9	30	7
small malignant (rad)	39	43	–	3	0	0	4
metastatic (rad)	160	178	–	3	3	9	2
benign	3,658	3,018	164	174	56	129	59
unknown malignancy	1,999	1,845	59	111	67	110	12
malignant (path)	288	390	79	–	–	–	26
small malignant (path)	5	13	8	–	–	–	4
primary malignant (path)	253	355	79	–	–	–	26
esophagus tumors							
total	2,720	3,141	388	40	189	95	78
small	217	290	64	5	45	2	32
malignant (rad)	722	789	–	14	54	28	14
small malignant (rad)	75	86	–	2	16	1	4
metastatic (rad)	70	91	–	0	9	8	0
benign	253	243	13	10	12	4	15
unknown malignancy	775	839	134	16	123	63	20
malignant (path)	970	1,270	235	–	–	–	29
small malignant (path)	56	110	42	–	–	–	24
primary malignant (path)	953	1,238	221	–	–	–	29
spleen tumors							
total	14,050	13,223	474	494	357	691	95
small	5,050	4,688	157	169	135	113	53
malignant (rad)	2,136	2,083	–	69	104	91	20
small malignant (rad)	851	827	–	20	36	2	15
metastatic (rad)	1,841	1,829	–	43	97	90	15
benign	6,283	5,779	155	88	32	176	26
unknown malignancy	5,582	5,278	283	337	221	424	47
malignant (path)	49	83	34	–	–	–	2
small malignant (path)	12	23	11	–	–	–	0
primary malignant (path)	6	6	0	–	–	–	0
adrenal gland tumors							
total	11,955	11,477	667	801	357	1,417	127
small	5,232	4,938	282	388	113	137	88
malignant (rad)	2,857	2,877	–	94	59	430	15
small malignant (rad)	977	983	–	16	19	6	11
metastatic (rad)	2,667	2,815	–	78	57	421	34
benign	3,387	3,142	166	73	55	343	51
unknown malignancy	5,430	5,018	363	634	243	644	37
malignant (path)	281	440	133	–	–	–	24
small malignant (path)	65	103	32	–	–	–	18
primary malignant (path)	74	74	0	–	–	–	0

Appendix B Comparison of our Reader Study to the Literature

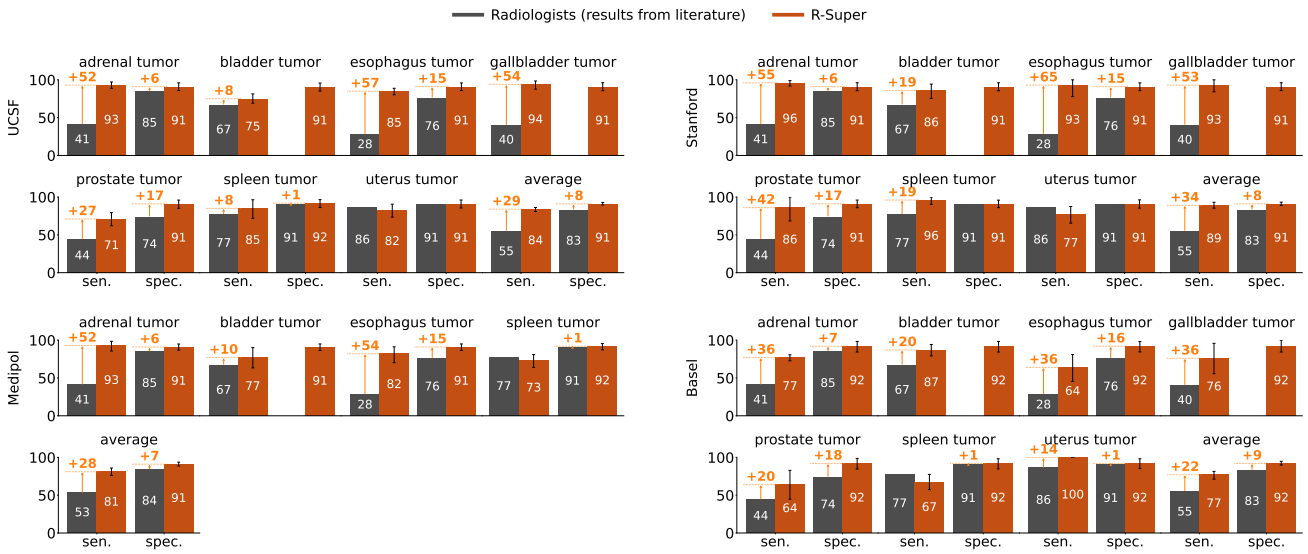

Fig. B1: **R-Super surpasses radiologist performances reported in the literature.** As a complement to our reader study, here we compare R-Super† to radiologist performances reported in previous studies [57–63]. We evaluate R-Super† on all our 4 test datasets for malignant tumor detection (malignant vs no-tumor), matching the task the radiologists performed in the previous studies [57–63]. The R-Super† model evaluated on Stanford Test was trained on UCSF Train, and the R-Super† model evaluated on the other test sets was trained on R-Super Train. We trained with all available reports and masks. Dataset details are in Tab. 1. We evaluate R-Super† at 91% specificity (or above), matching the maximum reported radiologist specificity. The sensitivity of R-Super† surpassed radiologists on average and for 5 tumor types (adrenal, bladder, esophagus, gallbladder, and prostate) across all datasets. Results for uterus and spleen were less conclusive: R-Super† surpassed radiologists on some datasets but not others, likely reflecting differences in dataset difficulty (e.g., tumor size distributions). Notably, this comparison of R-Super† and the literature is limited and just a complement to the reader study in Sec. 2.4, since here the radiologists and R-Super† were evaluated on different datasets, and in Sec. 2.4 they were evaluated on the same dataset. For bladder and gallbladder tumors the selected studies did not present the radiologist specificity. Error bars represent 95% confidence intervals, calculated with bootstrapping (1,000 bootstrap samples). We skipped evaluation on organs with less than 10 positive scans in the dataset (Medipol).

In this section, we demonstrate that the performances of radiologists in our reader study are consistent with results in the literature, and that R-Super can also surpass the radiologist performances in the literature. The radiologists' tumor detection sensitivity is not high for many tumor types studied here (Tab. D41), because these tumor types are very difficult to detect in CT scans. Accordingly, CT is not the primary diagnostic tool for tumors in the bladder, esophagus, prostate, and uterus. However, more than 300 million CT scans are performed yearly in the world, for diverse reasons. Inevitably, this large number of scans capture many early tumors, which often remain undetected [5]. AI has the potential to help radiologists find possible tumor cases that would have been missed. Then, such cases can be confirmed through specific, more accurate exams, which are usually not used for routine screening (e.g., cystoscopy).

We found in the literature a set of studies [57–63] where radiologists analyzed a dataset of patients with malignant tumors and normal patients (details in B.1). We extracted from these studies the sensitivity and specificity reported for the radiologists. For the bladder and gallbladder, the studies did not report specificity. Figure B1 compares the radiologist results in the literature to the performance of R-Super. Since the datasets used in the selected studies are not available, the figure shows the performance of R-Super in all our test datasets: UCSF as the internal test set, and Medipol, Stanford, and Basel as the external. We evaluate R-Super for malignant tumor detection (malignant vs no-tumor), the same task performed by the radiologists in the selected studies. We evaluate R-Super at 91% specificity, the highest radiologist specificity among the selected studies.

At matched or superior specificity, R-Super surpassed the radiologists' sensitivities reported in the literature on average and for 5 tumor types (adrenal, bladder, esophagus, gallbladder, and prostate) across all our test datasets (Fig. B1). For uterus and spleen tumors, results were mixed: R-Super surpassed the reported radiologist performance on some datasets but fell below it on others, likely reflecting differences in dataset difficulty such as tumor size distributions. This comparison is less controlled and more limited than our reader study (Sec. 2.4), since here the radiologists and R-Super were evaluated on different datasets, but in our reader study they were evaluated on the same datasets. Nevertheless, these results further demonstrate that R-Super can surpass radiologist performance, even when compared to independent results from the literature.

In our reader study, the radiologists achieved higher specificity than reported in the literature for all tumor types except uterus (89% versus 91% in [62], but within their reported 95% CI of 72-97%). For all other tumor types, specificity in the literature ranged from 74 to 91% (Fig. B1), while our radiologists achieved 96 to 99% (Fig. 5). This difference likely reflects a prevalence expectation effect [64, 65]. Our radiologists searched for tumors across multiple organs simultaneously (Supp. H), limiting the prevalence of any single tumor type in our reader study dataset (Tab. 1). In contrast, most literature studies focused on a single tumor type, often with high disease prevalence (e.g., 52% in [58], and 50% in [62]). Higher prevalence expectation is known to reduce specificity and increase sensitivity, as radiologists can unconsciously lower their diagnostic threshold, increasing false positives [64, 65]. In terms of sensitivity, our radiologists surpassed the literature in detecting esophagus and gallbladder tumors, had a similar result in prostate tumors, and were below the literature in spleen, adrenal, uterus, and bladder tumors. These differences can be explained by different tumor sizes in our dataset and datasets used in the literature, our dataset had a large proportion (47%) of small tumors, which are difficult to detect (Tab. A1). The differences can also be explained by the prevalence expectation effect, which can increase sensitivity in studies that focus on a single tumor type and have higher prevalence of that tumor.

B.1 Details of Selected Studies

A limitation of this section's comparison between the literature and R-Super is that they were evaluated in different test sets, with different patient populations, different CT scanners, possibly different proportions of the types of malignant tumors, different contrast protocols and different hospitals. Addressing this limitation is a key reason why we performed the reader study in Sec. 2.4.

Here, we provide a brief summary of each study.

Bladder tumors [57]. CT scans were selected for patients later diagnosed with bladder cancer (99 patients; 226 CTs). These scans were acquired up to 5 years before the pathologic diagnosis (pre-diagnostic). Radiologists achieved 67% tumor detection sensitivity. The study did not report specificity. Since these CT scans are pre-diagnostic, some may have a very small tumor, or truly no tumor, reducing the reported radiologist sensitivity.

Esophagus tumors [58]. Non-contrast CT scans come form 52 esophagus cancer patients and 48 normal patients. Radiologists achieved 25–31% sensitivity at 74–78% specificity. Unlike this study, our test dataset used contrast-enhanced CT (easier).

Gallbladder tumors [59]. This study, published in 1997, is a retrospective analysis. It covers gallbladder carcinoma patients at the Howard University, for the previous 27 years. Radiologist performance for gallbladder tumor detection in CT was reported as 40% sensitivity. Specificity is not reported.

Prostate tumors [60]. The study included 139 clinically significant prostate cancer CTs and 432 healthy CTs. Radiologists achieved 44% sensitivity and 74% specificity in detecting these tumors.

Spleen tumors [61]. A 2024 meta-analysis synthesized spleen tumor detection performance across different imaging modalities. On CT, radiologists achieved 74% sensitivity and 89% specificity in detecting spleen tumors.

Uterus tumors [62]. In asymptomatic postmenopausal women on CT (22 cancers; 22 controls), the endometrium thickness was measured by radiologists to detect endometrial cancer. An 8 mm thickness threshold yielded 86% sensitivity and 91% specificity for detecting the tumors. This study considers only endometrial cancers, but our test dataset may include other types of uterus malignant tumors.

Adrenal tumors [63]. The study considered 91 lung cancer patients. Of them, 53 had adrenal metastases (autopsy-validated). Radiologists could detect these metastasis on CT with sensitivity of 20 to 41%, but high specificity (84 to 99%).

Appendix C Supplementary Methods

C.1 Tumor Detection and Malignancy Identification

R-Super is a training framework for tumor segmentation. However, tumor detection and malignancy identification are essential tasks. These are classification tasks that determine whether each organ has tumors or not, and whether each organ has malignant tumors or not. Here, we perform this classification based on segmentation, which enables better tumor detection performance and interpretability (as seen in Sec. 2.2 and [12]). Therefore, we use classification heads to convert tumor segmentation outputs into classification outputs (Fig. C2).

The input of the classification heads is the output of the segmentation model. For each organ, the classification head receives the 3 channels of segmentation logits: tumor segmentation, malignant tumor segmentation, and benign tumor segmentation. Based on these channels the classification head outputs, for each organ, the probability of tumors, benign tumors, and malignant tumors. Because the classification head only needs to analyze the tumor segmentations, not the CT scan itself, it can be lightweight. Specifically, it uses three 3D convolutional blocks (each with a $3 \times 3 \times 3$ convolution and down-sampling by $2 \times 2 \times 2$), followed by 2 light transformer layers for better use of contextual information, and a final fully-connected layer with sigmoid activation. The heads are

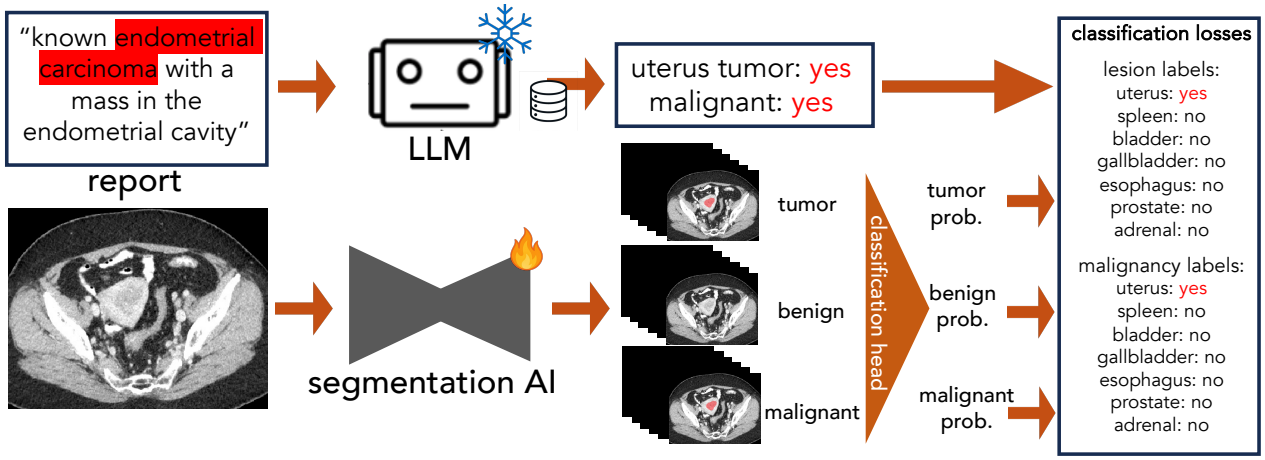

Fig. C2: **R-Super performs tumor detection and malignancy identification based on segmentation.** This approach substantially outperformed a standard classification model (without segmentation) in our experiments (Fig. 4). Tumor detection and malignancy identification are essentially classification tasks. R-Super solves them by applying a light classification head to the outputs of the segmentation model trained with the R-Super losses. The segmentation model has 3 channels per tumor type (tumor, benign, malignant). Therefore, the classification head outputs are a simple summary of the segmentation outputs. The classification head is trained jointly with the segmentation model, using a standard classification loss (cross-entropy) and labels extracted from reports.

trained with a standard classification loss (cross-entropy), using classification labels (per organ: tumor present/absent; malignant tumor present/absent; benign tumor present/absent) previously extracted from radiology and pathology reports by the LLM.

C.2 R-Super-guided Mask Creation

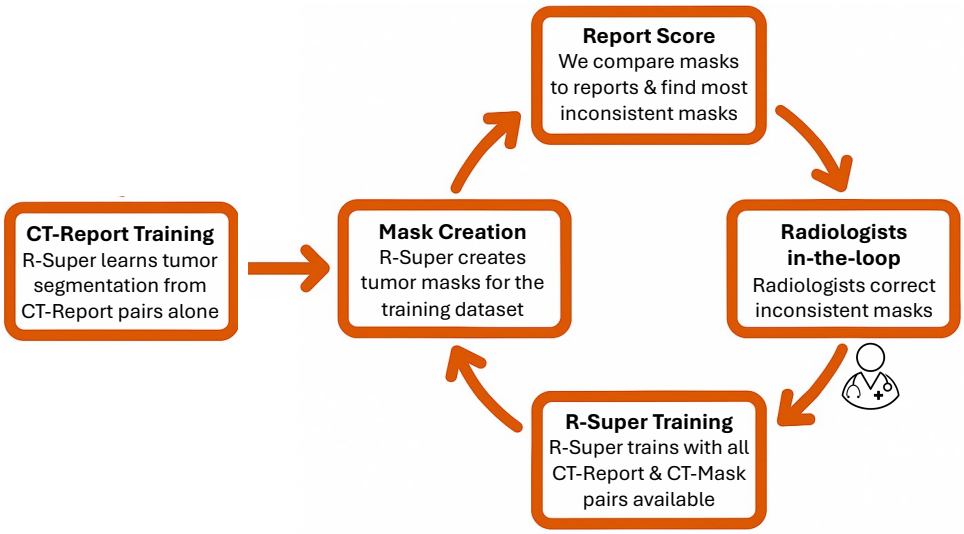

Fig. C3: **Our report-guided active learning enables radiologists to create tumor masks 3 to 5 times faster.** Instead of drawing masks from scratch, radiologists correct AI-generated masks and have access to the original radiology reports—reducing annotation time from 16 to 3 minutes for small tumors, and from 52 to 18 minutes for large tumors. The process prioritizes the most inaccurate masks, identified automatically using the Ball Loss, which increases when AI predictions disagree with report details (tumor number, size, or rough location). Retraining on these corrected masks helps R-Super avoid previous errors. Standard active learning begins with radiologists creating masks from scratch, without AI assistance—the cold-start problem. In contrast, our strategy starts with R-Super, already trained from CT–Report pairs, providing AI assistance from the start. As radiologist-corrected masks are created, R-Super continuously improves and consistently outperforms standard segmentation models (trained with masks, no reports), making it a more effective assistant throughout the entire active learning process.

Drawing a single tumor mask in CT (tri-dimensional) can take a radiologist over 30 minutes, making large-scale mask creation costly and slow [66, 67]. Our datasets include tumor masks created by 31 radiologists using a new report-guided active learning strategy based on R-Super (dataset summary in Tab. 1).

Figure C3 illustrates our active learning strategy. We first train R-Super on CT–Report pairs only. Then, we iteratively create tumor masks through a loop: (I) R-Super generates AI-made tumor masks; (II) we identify those the least consistent with reports; (III) radiologists correct them (the 200 most inconsistent masks); and (IV) we re-train R-Super with the radiologist-corrected tumor masks plus the remaining CT-Report pairs. Consistency between AI-made tumor masks and reports (step II) is quantified using the Ball Loss (Section 4.4), which increases when the AI-made tumor mask disagrees with tumor descriptions in reports—in tumor number, size, and rough location. Retraining on corrected masks (plus all CT-Report pairs) provides per-voxel guidance for difficult cases, and teaches R-Super to avoid its past mistakes.

Unlike traditional active learning (where radiologists begin by creating masks without AI assistance) our strategy starts with a strong segmentation model to assist radiologists, R-Super trained from CT–Report pairs alone (no mask). This provides effective AI assistance from the start, accelerating early mask creation. As more radiologist-corrected masks are added to the training dataset, R-Super continuously improves (Fig. 3). Furthermore, R-Super surpasses standard segmentation models (trained with masks and no reports), whether few or many masks are available (Sections 2.1, 2.2). Thus, R-Super provides better assistance throughout the whole active learning process.

We measured annotation time on 20 randomly selected tumor scans covering all seven tumor types, from the Stanford Test dataset. Each case was annotated twice by the same radiologist: once with R-Super assistance and once without. To minimize human learning effects, cases were alternately assigned to the AI-assisted and unassisted conditions in the first round, and the assignment was reversed in the second round. For scans containing only small tumors ($\leq$ 2 cm), annotation took an average of 3 minutes with R-Super assistance and 16 minutes without assistance. For scans containing larger tumors (>2 cm), annotation took 18 minutes with R-Super assistance and 52 minutes without assistance.

C.3 Volume Loss for Reports without Tumor Size or Count

When a report mentions that an organ has tumors, but it does not inform tumor diameters or the number of tumors in the organ, we cannot calculate the standard Volume Loss. That is because tumor numbers and diameters are needed to estimate the total tumor volume from the report ($V_{r,o}$). When this information is missing, we resort to a prior-based, high-tolerance version of the Volume Loss. This loss version only constrains the AI to segment tumors within a reasonable volume range, [V_{min},V_{max}]. I.e., we do not penalize the AI when it segments tumors with volumes within this range, but we penalize it when volumes are outside the range.

To define V_{min} and V_{max}, we analyzed the reports in our dataset (R-Super train): less than 3% of tumors have diameters smaller than 5 mm, and less than 2% have diameters larger than 120 mm. Therefore, we set $V_{min} = 65$ and $V_{max} = 904,779$, the volumes of balls with 5 mm and 120 mm of diameter, respectively. The high-tolerance Volume Loss is defined in Eq. C1. It is zero and has zero gradient when the segmented tumor volume ($V_{s,o}$) is in the range [V_{min},V_{max}].

$$L_{\text{vol},o} = L_{\text{forg},o}\left(V_{s,o},\, \widehat{V}_{r,o}\right) + L_{\text{bkg},o}(\mathbf{T}^o) \tag{C1}$$

$$\text{where} \quad \widehat{V}_{r,o} = \begin{cases} V_{min} & \text{if } V_{s,o} < V_{min}, \\ V_{s,o} & \text{if } V_{min} \leq V_{s,o} \leq V_{max}, \\ V_{max} & \text{if } V_{s,o} > V_{max}, \end{cases} \tag{C2}$$

C.4 Ball Loss for Reports without Tumor Size or Count

For reports missing tumor sizes, we use a prior-based version of the Ball Loss. As in the prior-based Volume Loss, we assume that the tumor has at least 5 mm in diameter, because less than 3% of tumors reported in our dataset have less than 5 mm. We first apply the Ball Convolution considering a tumor diameter of 5 mm. This locates the small tumor, possibly near the center of a real, larger tumor. As in the original Ball Loss, we maximize the probabilities of the located tumor, using cross-entropy and Dice losses. However, in the original Ball Loss, after locating all tumors mentioned in the report, we minimized the probabilities for all voxels not assigned to any of the located tumors. Here, we do not do not do this. We only maximize probabilities in voxels assigned to the located tumor, and we minimize probabilities outside the organ. Voxels inside the organ but not assigned to any located tumor are not optimized at all. The reason is that we do not know the real tumor size. Thus, it is not possible to estimate how many voxels inside the organ o are tumor voxels. We just enforce that, at least, a 5 mm tumor is segmented, but we do not penalize the segmentation model if it segments a larger tumor. In case a report does not inform how many tumors an organ has, we use the Ball Loss to localize and maximize probabilities for each tumor described in the report. However, we again do not penalize the voxels inside the organ but not assigned to any tumor. Thus, we do not penalize the model for segmenting more tumors.

C.5 Details on the Ball Loss

Weighted cross-entropy compensating for uncertain tumor borders. To better account for uncertainty in the tumor diameters in the reports, our cross-entropy function (CE in Eq. 19) gives higher weights ($w(x, y, z)$) to voxels where the AI predicts higher tumor probabilities, reducing the penalty on uncertain tumor borders (Eq. C3). To assign these weights, we rank the voxels in generated tumor mask $\mathbf{M}$ by predicted tumor probability in descending order, assigning rank $r(x, y, z) \in \{0, \ldots, n-1\}$, where n is the number of tumor voxels in $\mathbf{M}$ ($|\mathbf{M}|$) and $r = 0$ marks the highest-probability voxel. The weight $w(x, y, z)$ decays with the rank, as shown in Eq. C4; $c \in (0, 1)$ sets the decay sharpness (we set $c = 0.5$, where the most confident half of $\mathbf{M}$ carries half of the loss weight). The normalization in Eq. C4 ensures $\sum_{(x,y,z)\in\mathbf{M}} w(x, y, z) = n$, preserving the overall scale of the cross-entropy loss. Outside $\mathbf{M}$, the weight is set to $w = 1$ (standard cross-entropy).

$$\mathrm{CE}\big((1-\mathbf{M}')\odot\mathbf{T}^o,\,\mathbf{M}\big) = -\frac{1}{|\Omega|}\sum_{(x,y,z)\in\Omega} w(x,y,z)\left[\mathbf{M}\log\big((1-\mathbf{M}')\mathbf{T}^o\big) + (1-\mathbf{M})\log\big(1-(1-\mathbf{M}')\mathbf{T}^o\big)\right] \tag{C3}$$

$$w(x,y,z) = \begin{cases} \dfrac{n\,(1-c)^{r(x,y,z)/n}}{\sum_{r'=0}^{n-1}(1-c)^{r'/n}}, & \text{if } \mathbf{M}(x,y,z) = 1 \\ 1, & \text{otherwise} \end{cases} \tag{C4}$$

Improving tumor localization in the Ball convolution. To improve tumor localization, the Ball Convolution kernel has slightly lower values near the ball border (small Gaussian decay). This improves the alignment between the kernel center and the tumor center, where predicted tumor probabilities tend to be higher. However, this decay is mild, so that the Ball Convolution output is close to the sum of tumor probabilities inside each ball location. The Gaussian Ball Convolution kernel is shown in Eq. C5. In the equation, (u, v, w) are kernel coordinates, with origin at the kernel center, and d_i is tumor i's reported diameter. We set the Gaussian standard deviation σ_i to the ball radius multiplied by a constant α, keeping the decay identical across tumor sizes. We use $\alpha = 3$, so that the kernel falls only to about 0.95 of its center value at the boundary, ensuring a mild decay.

$$\mathbf{k}_i(u,v,w) = \begin{cases} e^{-\frac{u^2+v^2+w^2}{2\sigma_i^2}}, & \text{if } \sqrt{u^2+v^2+w^2} \le \dfrac{d}{2} \\ 0, & \text{otherwise} \end{cases} \tag{C5}$$

$$\text{where } \sigma_i = \alpha\,\frac{d_i}{2} \tag{C6}$$

Appendix D Supplementary Results

D.1 R-Super Results: Summary Tables

In this section, we summarize the performance of R-Super trained with only CT-Report pairs and R-Super† trained on CT-Report plus CT-Mask pairs. The tables in this section include aggregate results over all test sets, for all tumor types. First, we evaluate tumor detection, malignant tumor detection, and malignancy identification, using sensitivity, specificity, F1-Score and AUC. Next, we present tumor localization performance on the Stanford external test set (which includes a large number of tumor segmentation masks, 602). Localization performance is displayed in plots, which show both DSC and localization adjusted sensitivity, i.e., percentage of AI-made tumor segmentation masks that overlap with radiologist-made tumor segmentation masks. Comparisons between R-Super and public AI models are shown in Supp. D.6, comparisons to other training methods trained on our dataset in Supp. D.7, and comparisons to radiologists in Supp. D.8.

Table D2: **Tumor detection: R-Super detects tumors across multiple organs.** The results show tumor detection, i.e., identifying scans that contain tumors (benign or malignant) in a given organ versus no tumor scans. We show results for R-Super trained with CT-Report pairs only, and R-Super† trained with CT-Report plus CT-Mask pairs. The models tested on UCSF (internal, USA), Basel (external, Switzerland), and Medipol (external, Turkey) were trained on the R-Super train set (USA); models tested on Stanford (external, USA) were trained on UCSF train (USA). See train/test dataset details in Tab. A1. Results are shown as average and 95% CI (bootstrap, 1,000 samples). We show the operating point of best balanced accuracy.

AI	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	77 [72,81]	73 [64,80]	83 [79,86]	79 [74,83]	85 [81,88]	87 [80,93]	90 [87,92]	91 [88,94]
R-Super†	64 [59,69]	91 [85,96]	76 [73,80]	82 [78,86]	82 [79,86]	90 [85,95]	89 [86,91]	90 [88,93]
external evaluation: Stanford (USA)								
R-Super	75 [69,81]	82 [73,88]	81 [76,85]	82 [77,87]	76 [62,88]	69 [61,78]	58 [46,69]	79 [70,87]
R-Super†	72 [66,79]	84 [77,90]	80 [75,84]	84 [79,88]	78 [65,91]	92 [87,96]	78 [67,87]	90 [83,95]
external evaluation: Medipol (Türkiye)								
R-Super	90 [84,94]	65 [58,71]	77 [72,82]	84 [80,88]	77 [71,83]	89 [84,93]	82 [77,86]	89 [86,92]
R-Super†	61 [53,69]	94 [90,97]	72 [66,78]	87 [83,90]	83 [78,88]	80 [74,86]	82 [78,86]	89 [86,92]
external validation: Basel (Switzerland)								
R-Super	71 [65,76]	92 [84,98]	82 [78,86]	89 [84,92]	85 [77,92]	89 [80,96]	89 [83,93]	90 [84,94]
R-Super†	84 [80,89]	92 [84,98]	91 [88,93]	92 [88,94]	90 [83,96]	87 [77,95]	91 [86,95]	92 [87,96]

AI	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	73 [67,78]	72 [63,79]	79 [74,82]	78 [73,82]	83 [78,87]	87 [80,93]	88 [85,91]	89 [86,92]
R-Super†	78 [73,83]	80 [73,87]	83 [80,87]	84 [80,88]	80 [75,84]	90 [85,95]	87 [84,90]	88 [84,91]
external evaluation: Stanford (USA)								
R-Super	74 [69,80]	75 [67,83]	79 [75,84]	80 [75,85]	76 [71,80]	85 [78,91]	84 [81,87]	90 [87,93]
R-Super†	76 [70,81]	89 [83,94]	83 [79,87]	87 [83,91]	84 [80,87]	79 [71,86]	88 [85,90]	89 [84,92]
external evaluation: Medipol (Türkiye)								
R-Super	60 [52,67]	68 [62,75]	62 [56,68]	68 [63,74]	88 [82,93]	76 [70,82]	80 [74,84]	88 [84,92]
R-Super†	67 [60,74]	84 [78,89]	73 [68,78]	77 [72,82]	95 [90,98]	68 [62,76]	80 [75,84]	90 [87,93]
external validation: Basel (Switzerland)								
R-Super	70 [59,80]	53 [39,68]	69 [60,77]	64 [55,75]	64 [58,70]	100 [100,100]	78 [73,82]	83 [78,87]
R-Super†	55 [43,67]	98 [94,100]	71 [60,80]	84 [76,90]	81 [76,86]	98 [94,100]	89 [86,92]	90 [86,93]

AI	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	69 [63,76]	92 [87,97]	80 [74,84]	85 [80,90]	87 [85,90]	88 [81,93]	92 [90,94]	93 [90,95]
R-Super†	86 [81,90]	80 [71,87]	86 [82,90]	87 [82,91]	82 [79,85]	92 [87,97]	90 [88,91]	91 [89,93]
external evaluation: Stanford (USA)								
R-Super	76 [68,84]	90 [84,95]	82 [75,87]	89 [84,92]	84 [81,87]	91 [86,96]	91 [89,92]	93 [90,95]
R-Super†	94 [89,98]	68 [60,76]	82 [77,88]	90 [86,93]	83 [80,85]	90 [84,95]	90 [88,91]	90 [88,93]
external evaluation: Medipol (Türkiye)								
R-Super	80 [67,93]	97 [94,99]	81 [70,90]	93 [86,98]	88 [84,91]	91 [87,95]	92 [89,94]	94 [91,96]
R-Super†	89 [77,98]	97 [94,99]	86 [76,94]	96 [92,99]	82 [78,86]	92 [87,96]	88 [86,91]	90 [88,93]
external validation: Basel (Switzerland)								
R-Super	82 [76,88]	89 [79,96]	89 [84,92]	88 [83,93]	72 [69,74]	85 [74,94]	83 [82,85]	83 [78,88]
R-Super†	87 [82,92]	77 [66,88]	90 [86,93]	86 [81,91]	77 [75,79]	92 [84,98]	87 [86,88]	87 [83,90]

AI	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	76 [72,80]	92 [87,97]	86 [83,88]	88 [85,91]	78 [77,80]	84 [81,87]	85 [84,86]	86 [84,88]
R-Super†	75 [71,79]	95 [90,98]	85 [82,88]	87 [84,90]	78 [76,80]	88 [86,90]	85 [84,86]	87 [86,88]
external evaluation: Stanford (USA)								
R-Super	68 [64,72]	90 [85,95]	80 [77,83]	81 [77,84]	76 [73,78]	83 [81,86]	79 [77,81]	85 [83,86]
R-Super†	76 [73,80]	90 [83,95]	85 [83,88]	83 [80,86]	80 [78,83]	84 [82,87]	84 [82,86]	88 [86,89]
external evaluation: Medipol (Türkiye)								
R-Super	62 [56,67]	89 [84,94]	74 [70,78]	81 [77,84]	78 [75,80]	82 [80,84]	78 [76,81]	85 [84,87]
R-Super†	69 [65,74]	82 [76,88]	78 [74,81]	80 [76,84]	78 [75,81]	85 [83,88]	80 [78,82]	87 [86,89]
external validation: Basel (Switzerland)								
R-Super	67 [64,70]	85 [75,94]	80 [77,82]	82 [77,86]	73 [70,75]	85 [81,88]	81 [79,83]	83 [80,85]
R-Super†	77 [74,80]	81 [70,90]	86 [84,88]	82 [77,87]	79 [76,81]	90 [86,93]	86 [84,88]	88 [86,89]

Table D3: **Malignant tumor detection: R-Super detects malignant tumors across multiple organs.** The results show tumor detection stratified for malignant tumors (malignant vs no-tumor). We show results for R-Super trained with CT-Report pairs, and R-Super† trained with CT-Report plus CT-Mask pairs. The models tested on UCSF (internal, USA), Basel (external, Switzerland), and Medipol (external, Turkey) were trained on the R-Super train set (USA); models tested on Stanford (external, USA) were trained on UCSF train (USA). Train/test dataset details in Tab. A1. Results are shown as average and 95% CI (bootstrap, 1,000 samples). We show the operating point of best balanced accuracy. n/a: we skipped organs with less than 10 positive scans in the dataset.

AI	bladder malignant tumor				esophagus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	77 [71,82]	78 [70,85]	81 [76,85]	82 [77,86]	88 [83,92]	87 [80,93]	90 [87,93]	92 [89,95]
R-Super†	75 [69,81]	91 [85,96]	84 [79,87]	87 [83,91]	86 [82,90]	90 [85,95]	90 [87,93]	94 [91,96]
external evaluation: Stanford (USA)								
R-Super	82 [71,91]	83 [76,89]	76 [66,84]	84 [76,91]	73 [50,93]	89 [82,94]	56 [36,74]	83 [68,94]
R-Super†	84 [74,93]	97 [94,100]	88 [82,95]	94 [88,98]	93 [78,100]	96 [93,99]	85 [67,97]	98 [95,100]
external evaluation: Medipol (Türkiye)								
R-Super	82 [70,94]	94 [90,97]	78 [67,87]	90 [83,96]	93 [85,98]	93 [88,96]	86 [78,92]	97 [93,99]
R-Super†	72 [57,86]	99 [98,100]	82 [71,91]	93 [87,97]	96 [90,100]	80 [74,86]	73 [65,81]	95 [92,98]
external validation: Basel (Switzerland)								
R-Super	77 [68,86]	96 [90,100]	86 [80,92]	93 [88,97]	79 [61,92]	89 [80,96]	79 [64,88]	82 [69,92]
R-Super†	87 [79,94]	92 [84,98]	91 [85,95]	94 [89,97]	93 [82,100]	83 [72,92]	82 [71,92]	89 [79,97]

AI	gallbladder malignant tumor				uterus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	90 [82,96]	93 [88,97]	90 [85,94]	95 [91,99]	80 [71,88]	96 [92,99]	86 [80,91]	91 [86,95]
R-Super†	91 [84,96]	96 [93,99]	93 [88,97]	98 [97,99]	81 [72,90]	94 [89,98]	86 [80,91]	92 [87,96]
external evaluation: Stanford (USA)								
R-Super	76 [62,89]	96 [92,99]	80 [70,90]	92 [86,97]	81 [70,91]	93 [88,97]	83 [74,90]	95 [92,98]
R-Super†	93 [84,100]	96 [92,99]	90 [83,96]	96 [91,100]	88 [78,96]	86 [80,92]	81 [73,88]	92 [87,96]
external evaluation: Medipol (Türkiye)								
R-Super	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
R-Super†	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
external validation: Basel (Switzerland)								
R-Super	65 [40,90]	76 [64,87]	54 [32,71]	69 [54,84]	93 [83,100]	98 [94,100]	95 [88,100]	95 [87,100]
R-Super†	94 [80,100]	87 [77,96]	80 [63,92]	95 [89,99]	100 [100,100]	98 [94,100]	98 [95,100]	100 [100,100]

AI	prostate malignant tumor				adrenal malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	91 [85,96]	72 [63,80]	80 [73,86]	88 [82,93]	92 [87,96]	93 [88,97]	93 [90,96]	96 [94,98]
R-Super†	93 [88,98]	80 [71,87]	85 [79,90]	90 [85,94]	92 [88,96]	94 [89,98]	94 [90,96]	96 [94,99]
external evaluation: Stanford (USA)								
R-Super	90 [76,100]	90 [84,95]	74 [60,86]	95 [89,99]	95 [90,99]	96 [92,99]	95 [91,98]	98 [95,100]
R-Super†	86 [68,100]	94 [89,98]	78 [64,90]	97 [93,100]	98 [95,100]	90 [84,95]	93 [89,97]	98 [95,99]
external evaluation: Medipol (Türkiye)								
R-Super	n/a	n/a	n/a	n/a	95 [89,100]	91 [87,95]	86 [78,92]	97 [95,99]
R-Super†	n/a	n/a	n/a	n/a	92 [84,98]	95 [91,98]	88 [82,94]	96 [92,99]
external validation: Basel (Switzerland)								
R-Super	82 [67,96]	89 [79,96]	81 [67,90]	87 [77,95]	72 [68,76]	85 [74,94]	83 [80,86]	83 [78,88]
R-Super†	86 [72,97]	79 [68,89]	76 [62,86]	86 [74,94]	77 [73,81]	92 [84,98]	86 [84,89]	86 [82,90]

AI	spleen malignant tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	91 [81,100]	92 [87,97]	84 [74,92]	95 [90,98]	87 [84,90]	87 [84,90]	86 [84,88]	91 [89,93]
R-Super†	85 [72,96]	96 [91,99]	85 [74,93]	93 [86,98]	86 [84,89]	92 [90,93]	88 [86,90]	93 [91,94]
external evaluation: Stanford (USA)								
R-Super	90 [83,96]	90 [84,95]	87 [80,92]	92 [88,96]	84 [79,88]	91 [89,93]	79 [74,83]	91 [89,94]
R-Super†	96 [90,100]	95 [90,98]	94 [89,97]	96 [93,99]	91 [87,94]	93 [92,95]	87 [83,90]	96 [94,97]
external evaluation: Medipol (Türkiye)								
R-Super	78 [69,85]	89 [84,94]	79 [73,85]	87 [82,92]	87 [82,91]	92 [89,94]	82 [77,86]	93 [90,95]
R-Super†	81 [73,88]	89 [85,94]	81 [75,87]	90 [86,94]	85 [80,89]	91 [89,93]	81 [77,85]	94 [91,95]
external validation: Basel (Switzerland)								
R-Super	77 [67,86]	76 [63,86]	80 [73,86]	81 [74,88]	78 [73,83]	87 [83,90]	80 [75,83]	84 [81,87]
R-Super†	81 [73,89]	81 [70,90]	85 [78,90]	86 [79,92]	88 [85,91]	88 [84,91]	86 [81,89]	91 [88,93]

Table D4: **Small tumor detection.** The results show tumor detection (benign + malignant vs no-tumor), stratified to small tumors ($\leq$ 2 cm in diameter). We show results for R-Super trained with CT-Report pairs only, and R-Super† trained with CT-Report plus CT-Mask pairs. The models tested on UCSF (internal, USA), Basel (external, Switzerland), and Medipol (external, Turkey) were trained on the R-Super train set (USA); models tested on Stanford (external) were trained on UCSF train. Train/test dataset details in Tab. A1. Results are shown as average and 95% CI (bootstrap, 1,000 samples). We show the operating point of best balanced accuracy. n/a: we skipped organs with less than 10 positive scans in the dataset.

AI	bladder small tumor				esophagus small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	57 [48,66]	73 [64,80]	61 [54,69]	67 [60,74]	88 [79,95]	87 [80,93]	83 [75,89]	93 [89,96]
R-Super†	50 [40,59]	91 [85,96]	62 [53,71]	74 [67,81]	89 [81,97]	90 [85,95]	86 [79,92]	91 [84,97]
external evaluation: Stanford (USA)								
R-Super	73 [60,86]	82 [73,88]	65 [53,75]	81 [72,88]	n/a	n/a	n/a	n/a
R-Super†	80 [67,92]	84 [77,90]	72 [60,82]	84 [74,92]	n/a	n/a	n/a	n/a
external evaluation: Medipol (Türkiye)								
R-Super	91 [82,98]	65 [58,71]	59 [49,67]	80 [74,85]	86 [75,96]	89 [84,93]	74 [63,84]	93 [89,97]
R-Super†	37 [24,50]	94 [90,97]	47 [33,60]	80 [73,85]	89 [78,98]	80 [74,86]	66 [56,75]	90 [84,95]
external validation: Basel (Switzerland)								
R-Super	51 [36,68]	92 [84,98]	63 [49,77]	83 [74,91]	n/a	n/a	n/a	n/a
R-Super†	86 [75,97]	92 [84,98]	88 [79,94]	92 [84,98]	n/a	n/a	n/a	n/a

AI	gallbladder small tumor				uterus small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	55 [43,67]	72 [63,79]	54 [43,64]	70 [62,77]	75 [63,86]	87 [80,93]	75 [65,83]	86 [79,92]
R-Super†	62 [49,73]	80 [73,87]	62 [52,72]	76 [68,84]	63 [52,75]	90 [85,95]	69 [59,79]	81 [73,88]
external evaluation: Stanford (USA)								
R-Super	68 [54,82]	75 [67,83]	58 [45,68]	76 [66,85]	74 [61,87]	85 [78,91]	71 [59,80]	87 [80,93]
R-Super†	73 [58,86]	89 [83,94]	71 [59,82]	88 [82,94]	74 [63,86]	79 [71,86]	66 [55,76]	84 [78,90]
external evaluation: Medipol (Türkiye)								
R-Super	56 [45,67]	68 [62,75]	49 [39,57]	68 [61,76]	100 [100,100]	76 [70,82]	36 [20,50]	83 [76,88]
R-Super†	75 [65,84]	84 [78,89]	70 [62,78]	80 [73,87]	83 [57,100]	68 [62,76]	26 [12,38]	80 [72,88]
external validation: Basel (Switzerland)								
R-Super	n/a	n/a	n/a	n/a	42 [14,71]	100 [100,100]	59 [25,83]	78 [54,97]
R-Super†	n/a	n/a	n/a	n/a	67 [40,100]	98 [94,100]	76 [53,95]	90 [75,100]

AI	prostate small tumor				adrenal small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	42 [24,59]	92 [87,97]	49 [30,64]	72 [59,82]	88 [84,92]	88 [81,93]	91 [88,93]	93 [90,96]
R-Super†	77 [60,91]	80 [71,87]	62 [47,74]	81 [71,89]	81 [76,86]	92 [87,97]	88 [85,91]	90 [86,93]
external evaluation: Stanford (USA)								
R-Super	n/a	n/a	n/a	n/a	87 [83,90]	91 [86,96]	92 [89,94]	94 [92,96]
R-Super†	n/a	n/a	n/a	n/a	84 [80,87]	90 [84,95]	90 [87,92]	92 [89,94]
external evaluation: Medipol (Türkiye)								
R-Super	n/a	n/a	n/a	n/a	94 [89,98]	91 [87,95]	90 [86,94]	96 [93,98]
R-Super†	n/a	n/a	n/a	n/a	87 [80,92]	92 [87,96]	87 [82,91]	91 [87,95]
external validation: Basel (Switzerland)								
R-Super	69 [42,94]	89 [79,96]	64 [40,83]	82 [66,95]	65 [57,74]	85 [74,94]	76 [70,82]	79 [73,85]
R-Super†	85 [62,100]	77 [66,88]	61 [40,77]	84 [68,96]	70 [62,78]	92 [84,98]	81 [75,86]	84 [78,90]

AI	spleen small tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	77 [70,83]	92 [87,97]	84 [79,89]	86 [80,90]	69 [65,73]	84 [81,87]	71 [67,75]	81 [78,84]
R-Super†	75 [68,82]	95 [90,98]	84 [79,88]	86 [81,91]	71 [67,74]	88 [86,90]	73 [70,76]	83 [80,85]
external evaluation: Stanford (USA)								
R-Super	70 [63,77]	90 [85,95]	80 [74,84]	83 [79,88]	75 [70,80]	85 [82,87]	73 [68,77]	84 [81,87]
R-Super†	72 [65,79]	90 [83,95]	80 [76,85]	84 [79,88]	77 [72,82]	86 [83,89]	76 [72,80]	86 [84,89]
external evaluation: Medipol (Türkiye)								
R-Super	71 [63,78]	89 [84,94]	77 [70,82]	85 [81,89]	83 [80,86]	80 [77,82]	64 [60,68]	84 [82,86]
R-Super†	73 [66,81]	82 [76,88]	74 [68,80]	81 [76,86]	74 [69,79]	83 [80,86]	62 [57,66]	84 [81,86]
external validation: Basel (Switzerland)								
R-Super	60 [50,68]	85 [75,94]	71 [64,78]	75 [67,82]	58 [48,66]	90 [86,94]	67 [57,74]	79 [73,85]
R-Super†	62 [54,71]	81 [70,90]	73 [65,79]	76 [68,84]	74 [67,82]	88 [84,92]	76 [69,82]	85 [80,89]

Table D5: **Small malignant tumor detection.** The results show tumor detection stratified for small malignant tumors (malignant vs no-tumor). We show results for R-Super trained with CT-Report pairs, and R-Super trained with CT-Report plus CT-Mask pairs (R-Super†). The models tested on UCSF (internal, USA), Basel (external, Switzerland), and Medipol (external, Turkey) were trained on the R-Super train set (USA); models tested on Stanford (external, USA) were trained on UCSF train (USA). See train/test dataset details in Tab. A1. Results are shown as average and 95% CI (bootstrap, 1,000 samples). We show the operating point of best balanced accuracy. n/a: we skipped organs with less than 10 positive scans in the dataset.

AI	bladder small malignant tumor				esophagus small malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	58 [46,71]	78 [70,85]	57 [46,67]	70 [60,78]	95 [88,100]	87 [80,93]	82 [73,90]	96 [93,99]
R-Super†	69 [57,80]	91 [85,96]	74 [63,82]	82 [73,90]	100 [100,100]	90 [85,95]	88 [81,94]	99 [98,100]
external evaluation: Stanford (USA)								
R-Super	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
R-Super†	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
external evaluation: Medipol (Türkiye)								
R-Super	n/a	n/a	n/a	n/a	88 [70,100]	93 [88,96]	65 [46,81]	96 [87,100]
R-Super†	n/a	n/a	n/a	n/a	88 [68,100]	80 [74,86]	42 [28,56]	93 [85,99]
external validation: Basel (Switzerland)								
R-Super	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
R-Super†	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

AI	gallbladder small malignant tumor				uterus small malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
R-Super†	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
external evaluation: Stanford (USA)								
R-Super	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
R-Super†	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
external evaluation: Medipol (Türkiye)								
R-Super	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
R-Super†	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
external validation: Basel (Switzerland)								
R-Super	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
R-Super†	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

AI	prostate small malignant tumor				adrenal small malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	80 [58,100]	72 [63,80]	41 [23,56]	81 [65,93]	81 [68,94]	93 [88,97]	79 [66,88]	91 [83,98]
R-Super†	87 [67,100]	80 [71,87]	51 [33,67]	86 [74,96]	81 [67,94]	94 [89,98]	80 [68,90]	89 [79,97]
external evaluation: Stanford (USA)								
R-Super	n/a	n/a	n/a	n/a	88 [69,100]	96 [92,99]	80 [62,93]	93 [81,100]
R-Super†	n/a	n/a	n/a	n/a	94 [80,100]	90 [84,95]	70 [53,84]	93 [80,100]
external evaluation: Medipol (Türkiye)								
R-Super	n/a	n/a	n/a	n/a	90 [73,100]	91 [87,95]	65 [49,79]	96 [90,99]
R-Super†	n/a	n/a	n/a	n/a	90 [73,100]	95 [91,98]	76 [60,88]	93 [84,99]
external validation: Basel (Switzerland)								
R-Super	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
R-Super†	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

AI	spleen small malignant tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	82 [56,100]	92 [87,97]	62 [36,80]	94 [84,100]	79 [71,87]	84 [80,88]	64 [56,71]	86 [82,90]
R-Super†	91 [71,100]	96 [91,99]	77 [56,92]	94 [82,100]	86 [79,91]	90 [88,92]	74 [67,80]	90 [86,93]
external evaluation: Stanford (USA)								
R-Super	90 [76,100]	90 [84,95]	72 [57,85]	92 [82,98]	89 [77,98]	92 [89,96]	76 [65,85]	92 [84,98]
R-Super†	90 [76,100]	95 [90,98]	82 [67,92]	92 [81,100]	92 [82,100]	92 [89,95]	76 [65,85]	92 [84,99]
external evaluation: Medipol (Türkiye)								
R-Super	72 [56,86]	89 [84,94]	64 [51,76]	84 [75,92]	83 [74,91]	91 [88,94]	65 [55,73]	92 [87,95]
R-Super†	75 [60,88]	89 [85,94]	66 [53,77]	88 [80,94]	84 [75,92]	88 [85,91]	61 [53,69]	91 [86,95]
external validation: Basel (Switzerland)								
R-Super	n/a	n/a	n/a	n/a	32 [17,63]	87 [83,90]	15 [7,24]	55 [49,85]
R-Super†	n/a	n/a	n/a	n/a	58 [30,82]	88 [84,91]	18 [9,28]	77 [64,92]

Table D6: **Malignancy identification: R-Super differentiates between benign and malignant tumors.** The results shows malignancy identification, i.e., identifying malignant tumors versus benign tumors plus no-tumor cases. We show results for R-Super trained with CT-Report pairs, and R-Super trained with CT-Report plus CT-Mask pairs (R-Super†). The models tested on UCSF (internal, USA), Basel (external, Switzerland), and Medipol (external, Turkey) were trained on the R-Super train set (USA); models tested on Stanford (external, USA) were trained on UCSF train (USA). See train/test dataset details in Tab. A1. Results are shown as average and 95% CI (bootstrap, 1,000 samples). We show the operating point of best balanced accuracy. n/a: we skipped organs with less than 10 positive scans in the dataset.

AI	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	74 [68,80]	86 [80,91]	81 [76,85]	84 [80,88]	77 [72,82]	88 [82,94]	84 [80,88]	87 [83,91]
R-Super†	78 [71,83]	88 [82,93]	83 [78,87]	87 [82,90]	86 [82,90]	89 [83,94]	90 [87,93]	93 [91,95]
external evaluation: Stanford (USA)								
R-Super	78 [66,88]	83 [76,88]	70 [59,79]	80 [72,88]	73 [50,93]	86 [80,92]	51 [32,69]	81 [67,93]
R-Super†	84 [74,93]	92 [87,96]	81 [73,89]	92 [87,97]	93 [78,100]	93 [88,97]	74 [53,88]	95 [88,99]
external evaluation: Medipol (Türkiye)								
R-Super	85 [72,96]	91 [87,95]	72 [61,83]	92 [86,97]	82 [70,91]	92 [87,95]	77 [67,85]	88 [81,94]
R-Super†	72 [57,86]	98 [95,100]	78 [67,88]	91 [85,97]	82 [70,91]	92 [87,95]	77 [68,86]	94 [90,97]
external validation: Basel (Switzerland)								
R-Super	83 [75,91]	86 [77,94]	86 [80,92]	88 [82,94]	64 [46,82]	91 [84,98]	71 [53,83]	78 [66,89]
R-Super†	86 [77,94]	85 [75,94]	87 [80,92]	91 [85,95]	86 [70,97]	82 [72,92]	77 [63,88]	87 [76,94]

AI	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	85 [77,93]	82 [77,88]	77 [70,83]	88 [83,93]	80 [71,88]	79 [74,85]	70 [62,78]	83 [77,88]
R-Super†	91 [84,96]	89 [84,93]	85 [78,90]	95 [93,98]	80 [71,89]	75 [69,82]	68 [60,76]	82 [76,88]
external evaluation: Stanford (USA)								
R-Super	76 [62,89]	87 [82,92]	65 [52,76]	87 [81,92]	72 [60,84]	86 [81,90]	59 [48,68]	86 [80,91]
R-Super†	83 [71,93]	85 [80,90]	66 [54,76]	90 [83,95]	63 [50,74]	75 [70,80]	43 [34,52]	73 [66,80]
external evaluation: Medipol (Türkiye)								
R-Super	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
R-Super†	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
external validation: Basel (Switzerland)								
R-Super	71 [47,92]	69 [58,80]	49 [29,64]	66 [47,85]	60 [41,77]	93 [90,97]	60 [44,73]	83 [75,90]
R-Super†	94 [80,100]	89 [81,96]	80 [65,92]	95 [88,100]	93 [82,100]	51 [44,58]	38 [27,48]	75 [67,83]

AI	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	53 [42,64]	82 [76,88]	59 [49,68]	78 [72,84]	87 [81,93]	74 [68,80]	77 [71,82]	86 [82,90]
R-Super†	90 [83,96]	72 [64,80]	78 [71,84]	86 [81,91]	92 [87,96]	63 [57,70]	74 [69,79]	80 [76,85]
external evaluation: Stanford (USA)								
R-Super	67 [46,87]	99 [96,100]	76 [58,89]	92 [85,97]	93 [87,98]	70 [63,76]	73 [67,79]	88 [84,91]
R-Super†	81 [61,96]	90 [85,94]	64 [47,77]	94 [88,99]	78 [69,86]	86 [81,91]	76 [68,82]	89 [85,92]
external evaluation: Medipol (Türkiye)								
R-Super	n/a	n/a	n/a	n/a	90 [82,97]	74 [68,80]	62 [53,70]	86 [80,91]
R-Super†	n/a	n/a	n/a	n/a	88 [80,96]	73 [67,78]	60 [50,68]	81 [76,86]
external validation: Basel (Switzerland)								
R-Super	43 [24,62]	98 [94,100]	58 [36,74]	77 [64,88]	60 [55,64]	71 [66,75]	64 [60,68]	68 [64,71]
R-Super†	71 [52,86]	91 [82,98]	76 [60,86]	83 [71,93]	34 [30,38]	89 [86,92]	47 [42,52]	61 [57,65]

AI	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
internal evaluation: UCSF (USA)								
R-Super	79 [64,92]	75 [68,82]	54 [41,65]	80 [72,88]	76 [73,80]	81 [78,84]	72 [68,74]	84 [81,86]
R-Super†	85 [72,96]	77 [70,84]	58 [45,69]	84 [75,90]	86 [83,89]	79 [76,81]	76 [74,79]	87 [85,89]
external evaluation: Stanford (USA)								
R-Super	83 [73,91]	72 [66,79]	63 [54,70]	82 [76,88]	77 [72,82]	83 [81,85]	65 [60,69]	85 [82,88]
R-Super†	90 [82,97]	73 [67,79]	67 [59,74]	87 [83,92]	82 [77,86]	85 [82,87]	67 [62,71]	89 [86,91]
external evaluation: Medipol (Türkiye)								
R-Super	86 [79,92]	74 [68,80]	72 [65,78]	85 [80,89]	85 [80,90]	83 [80,85]	71 [66,75]	88 [85,90]
R-Super†	88 [80,93]	73 [67,79]	72 [65,78]	88 [83,91]	82 [77,87]	84 [81,86]	72 [67,76]	88 [86,91]
external validation: Basel (Switzerland)								
R-Super	68 [58,78]	64 [57,70]	53 [45,61]	68 [62,75]	64 [58,70]	82 [79,84]	63 [58,68]	76 [71,79]
R-Super†	66 [56,76]	69 [63,75]	55 [46,62]	70 [63,76]	76 [71,80]	80 [76,82]	66 [61,69]	80 [77,83]

D.2 Segmentation DSC and Inter-Reader Agreement

Table D7: **Inter-reader agreement is relatively low for the 7 tumor types we study, reflecting the difficulty of detecting and segmenting them.** Here, average inter-reader DSC was 56%, lower than values observed in some public datasets, such as the 86% inter-reader DSC observed for pancreatic tumors in PanTS [18]. This is expected since CT is often not the primary detection modality for these tumor types, unlike for pancreatic tumors. To quantify inter-reader variability, two radiologists from the same hospital, both with two years of experience, independently segmented tumors in 70 test CT scans (10 per tumor type), and we compared their masks using DSC, NSD, and HD95. Dataset size (small, large): adrenal (5, 5); bladder (5, 5); esophagus (3, 7); gallbladder (5, 5); prostate (5, 5); spleen (3, 7); uterus (2, 8).

metric	adrenal tumor	bladder tumor	esophagus tumor	gallbladder tumor	prostate tumor	spleen tumor	uterus tumor	average
small lesions								
DSC	50	43	46	44	47	55	52	48
NSD	55	40	40	45	48	57	50	48
HD95	15 mm	20 mm	22 mm	17 mm	15 mm	12 mm	13 mm	16 mm
large lesions								
DSC	66	53	52	59	60	70	64	61
NSD	69	56	51	57	60	75	65	62
HD95	9 mm	16 mm	19 mm	14 mm	12 mm	7 mm	10 mm	12 mm
all lesions								
DSC	58	48	50	52	53	66	62	56
NSD	62	48	48	51	54	70	62	56
HD95	12 mm	18 mm	20 mm	16 mm	14 mm	8 mm	11 mm	14 mm

The Dice score (DSC) assesses the overlap between tumor masks produced by a segmentation model and ground-truth tumor masks drawn by radiologists. The normalized surface distance (NSD) measures how well the segmented tumor border matches the tumor border in the radiologist-drawn mask. Tab. D8 reports both metrics for R-Super and all other segmentation models in this study, evaluated externally on the Stanford Test set, for which our radiologists created 602 masks. The table includes models trained on our UCSF train dataset (Tab. 1), as well as public segmentation models trained on ULS and FLARE.

R-Super†, trained on 116,514 reports and 870 masks, substantially and significantly ($p<0.05$) surpassed a segmentation model trained on the 870 masks alone, by +14% DSC (Tab. D8). R-Super† also surpassed every alternative training method trained on our dataset (CLIP, MTL, ModelsGenesis, nnU-Net) by at least +14% DSC, and surpassed the public tumor segmentation models by even larger margins, of at least +31% DSC (Tab. D8).

Furthermore, all baseline models trained on our datasets surpassed the public tumor segmentation models. This gap likely reflects the composition of public abdominal CT datasets, which are dominated by liver, kidney, and pancreas tumors and leave the tumor types we study strongly underrepresented. Our public data release fills this gap in the public domain, and R-Super is a highly effective framework for learning segmentation from such large-scale data.

Is DSC low? Inter-reader agreement provides a useful reference point for interpreting DSC. We asked two radiologists from the same hospital, both with 2 years of experience, to independently annotate the same 70 CT scans from our test dataset (randomly selected, 10 per tumor type). We then computed the inter-reader DSC and NSD between the two sets of radiologist-drawn masks (Tab. D7). The average inter-reader DSC was 56%, below the inter-reader DSC reported for other tumor types, such as pancreatic tumors in PanTS (86%) [18]. This is consistent with the nature of the tumor types studied here: they are known to be difficult to see and outline on CT, which is not the primary detection modality for several of these tumor types (e.g., bladder, prostate, and esophagus tumors). Even so, CT remains a critical modality for incidental tumor detection: over 300 million CT scans are performed worldwide each year for many clinical indications, inevitably capturing all types of tumor; AI could help radiologists flag these tumors, prompting timely confirmation with specialized exams such as endoscopy or MRI.

Importantly, in scans where R-Super† detected the tumor, its segmentation output overlapped the radiologist-drawn mask 86% of the time. This was measured on the radiologist-drawn masks from Stanford test (external). The FROC curves in Fig. 3b further confirm that the model is detecting and localizing the real tumors, even when contours are imperfect. The relatively low DSC therefore mainly reflects the lower inter-reader agreement and possible contouring inaccuracies, not critical failures of tumor detection or localization. Although contouring is important, tumor detection is the main goal of this study. Here, we demonstrate that detection through report-based segmentation is interpretable and accurate, as R-Super strongly surpassed a classification model and VLMs in tumor detection (Fig.2).

Table D8: **R-Super surpasses public segmentation models and alternative training methods in segmentation DSC, in external validation.** R-Super† trained with reports and masks significantly (p<0.05) surpassed the mask-based segmentation model by +14% DSC on average. The DSC values in the table are below 50%. This is expected, since even the DSC between two radiologists reached only 56% on average (Tab. D7), given the high difficulty of detecting and segmenting these tumor types in CT. Importantly, R-Super† produced tumor masks that overlapped with the radiologist-drawn mask on 86% of the test cases where the tumor was detected, indicating that R-Super† is detecting and localizing real tumors, even when contours and DSC are imperfect. Models were externally tested on Stanford Test for tumor segmentation. The models we trained were trained on the UCSF train (Tab. 1). The test set include the following radiologist-drawn tumor masks: bladder (95), gallbladder (76), esophagus (51), uterus (91), prostate (68), spleen (101), and adrenal glands (120). The NSD (normalized surface distance) tolerance is 3 mm. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model for each metric. We used paired permutation tests (two-sided, 10,000 permutations). Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples.

AI	bladder tumor		esophagus tumor	
	DSC	NSD	DSC	NSD
Public AI Models				
ULS	8 [5,12]	11 [8,15]	0 [0,0]	0 [0,1]
FLARE 23 Top1	0 [0,0]	0 [0,0]	0 [0,0]	0 [0,0]
FLARE 23 Top2	0 [0,0]	0 [0,0]	0 [0,1]	1 [0,2]
FLARE 24 Top1	1 [0,2]	1 [0,2]	0 [0,0]	0 [0,1]
FLARE 25 Baseline	2 [0,3]	2 [1,3]	0 [0,0]	0 [0,0]
AI Models Trained on UCSF				
CLIP	19 [15,24]	19 [14,24]	21 [14,27]	23 [16,29]
MTL	25 [20,30]	27 [21,32]	21 [15,28]	22 [16,27]
ModelsGenesis	20 [15,26]	19 [15,24]	25 [19,33]	28 [21,35]
segmentation	19 [14,24]	19 [14,24]	28 [21,35]	31 [25,38]
nnUNet	9 [5,13]	8 [4,12]	10 [4,18]	9 [3,16]
R-Super	8 [6,11]	8 [6,11]	6 [3,10]	8 [5,11]
R-Super†	34 [28,40]	34 [28,40]	23 [18,28]	25 [20,31]

AI	gallbladder tumor		uterus tumor	
	DSC	NSD	DSC	NSD
Public AI Models				
ULS	2 [1,5]	2 [1,4]	8 [5,12]	6 [3,9]
FLARE 23 Top1	7 [3,11]	6 [3,9]	2 [0,5]	2 [1,4]
FLARE 23 Top2	13 [7,19]	13 [8,19]	0 [0,0]	0 [0,0]
FLARE 24 Top1	9 [5,14]	7 [3,11]	9 [5,13]	5 [3,7]
FLARE 25 Baseline	6 [3,10]	5 [2,8]	16 [10,21]	10 [7,14]
AI Models Trained on UCSF				
CLIP	15 [11,19]	23 [17,29]	28 [22,35]	22 [17,26]
MTL	11 [7,15]	16 [11,21]	40 [33,47]	32 [27,38]
ModelsGenesis	15 [11,20]	21 [16,27]	29 [23,35]	23 [19,27]
segmentation	20 [15,25]	26 [20,33]	28 [22,35]	22 [18,27]
nnUNet	0 [0,0]	0 [0,0]	21 [15,29]	19 [13,25]
R-Super	19 [14,24]	20 [15,24]	21 [17,26]	17 [14,19]
R-Super†	41 [36,47]	44 [37,50]	42 [37,49]	32 [27,37]

AI	prostate tumor		adrenal tumor	
	DSC	NSD	DSC	NSD
Public AI Models				
ULS	26 [17,36]	17 [10,25]	0 [0,0]	0 [0,0]
FLARE 23 Top1	1 [0,2]	0 [0,1]	5 [2,8]	7 [3,10]
FLARE 23 Top2	0 [0,0]	0 [0,0]	7 [4,11]	10 [6,14]
FLARE 24 Top1	3 [0,7]	1 [0,3]	8 [4,12]	10 [5,15]
FLARE 25 Baseline	12 [5,19]	6 [3,10]	5 [2,8]	6 [3,10]
AI Models Trained on UCSF				
CLIP	15 [12,18]	9 [6,13]	22 [16,27]	29 [22,35]
MTL	12 [9,15]	9 [5,13]	32 [26,38]	44 [37,51]
ModelsGenesis	12 [8,16]	8 [4,12]	24 [19,29]	32 [26,38]
segmentation	17 [13,21]	12 [8,16]	33 [27,39]	43 [37,50]
nnUNet	8 [5,12]	6 [2,12]	24 [19,30]	30 [23,38]
R-Super	29 [22,36]	23 [17,29]	30 [27,34]	41 [37,45]
R-Super†	41 [33,48]	30 [24,37]	45 [40,50]	59 [54,64]

AI	spleen tumor		**Average**	
	DSC	NSD	DSC	NSD
Public AI Models				
ULS	3 [0,6]	3 [1,5]	7 [5,9]	6 [4,7]
FLARE 23 Top1	6 [3,10]	9 [5,14]	3 [2,4]	3 [2,4]
FLARE 23 Top2	13 [8,17]	16 [10,22]	5 [4,6]	6 [4,7]
FLARE 24 Top1	9 [4,13]	10 [5,14]	6 [4,7]	5 [4,6]
FLARE 25 Baseline	9 [4,14]	10 [5,15]	7 [6,9]	6 [4,7]
AI Models Trained on UCSF				
CLIP	24 [18,30]	29 [23,36]	21 [19,22]	22 [20,24]
MTL	28 [22,35]	34 [27,42]	24 [22,26]	26 [24,28]
ModelsGenesis	25 [19,31]	30 [23,36]	21 [19,23]	23 [21,25]
segmentation	28 [22,34]	31 [25,37]	24 [22,27]	26 [24,29]
nnUNet	18 [12,24]	20 [13,27]	13 [11,15]	13 [11,15]
R-Super	17 [13,21]	20 [16,24]	19 [17,21]	20 [18,21]
R-Super†	39 [34,45]	47 [40,53]	38 [36,40]	39 [36,41]

D.3 Scaling Performance

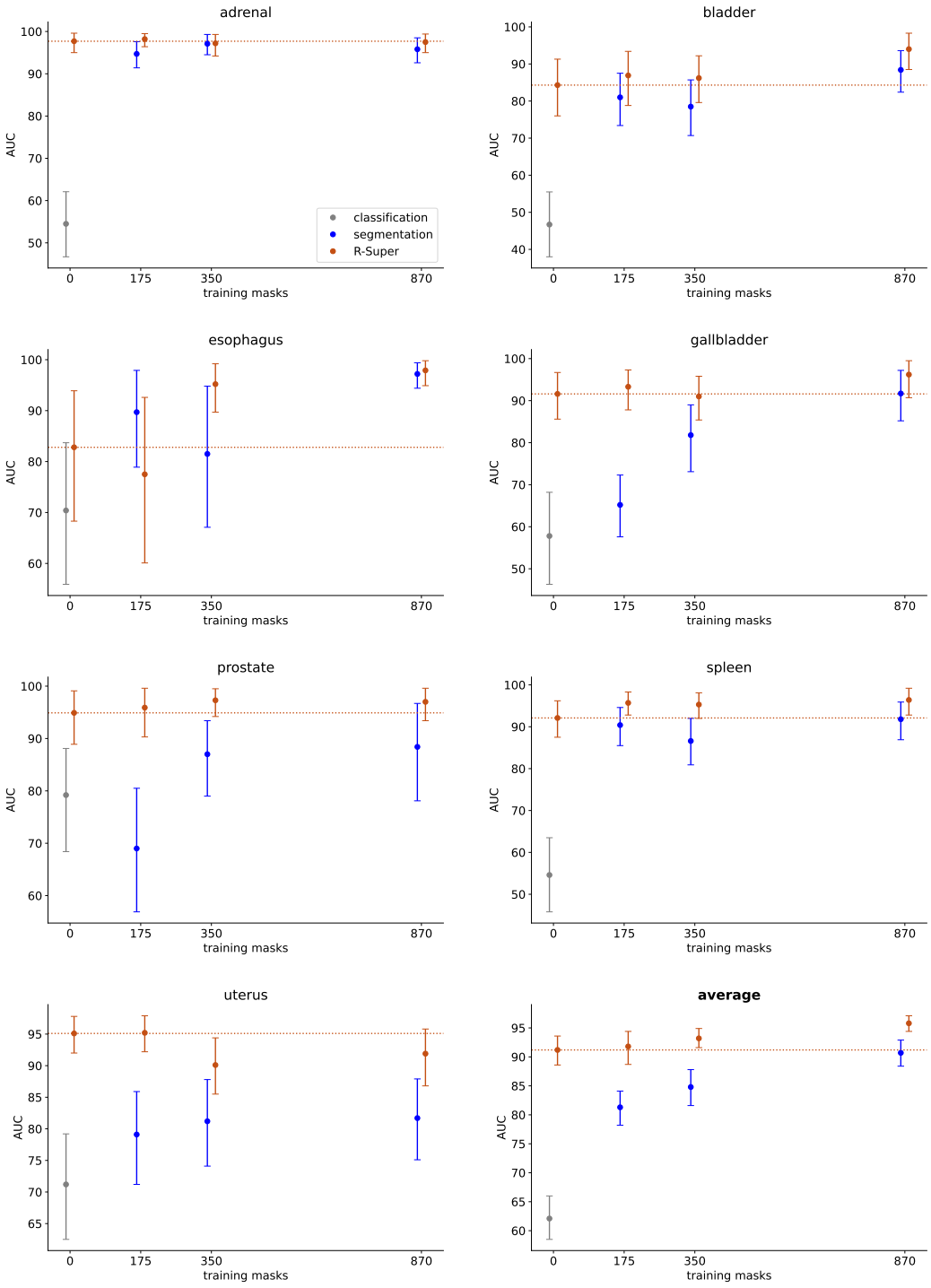

Fig. D4: Training with many reports surpasses training with fewer masks in tumor detection, and CT-Report plus CT-Mask training consistently surpasses mask-only training. The vertical axes show AUC for tumor detection stratified for malignant tumors. The horizontal axes indicate the number of training masks. On average across all tumor types, R-Super trained on reports alone (116,514 reports, no masks) substantially surpassed a segmentation model trained on 175 and 350 masks (no reports), and was slightly above the model trained on 870 masks; when trained on reports and masks, R-Super† significantly surpassed the segmentation model at all mask counts. R-Super (orange) was trained on reports only (leftmost point) or reports plus masks (other points); the segmentation model (blue) was trained on masks only. Both models were trained on the UCSF train set and tested on the whole Stanford test set. Error bars represent 95% confidence intervals, calculated with bootstrapping (1,000 bootstrap samples).

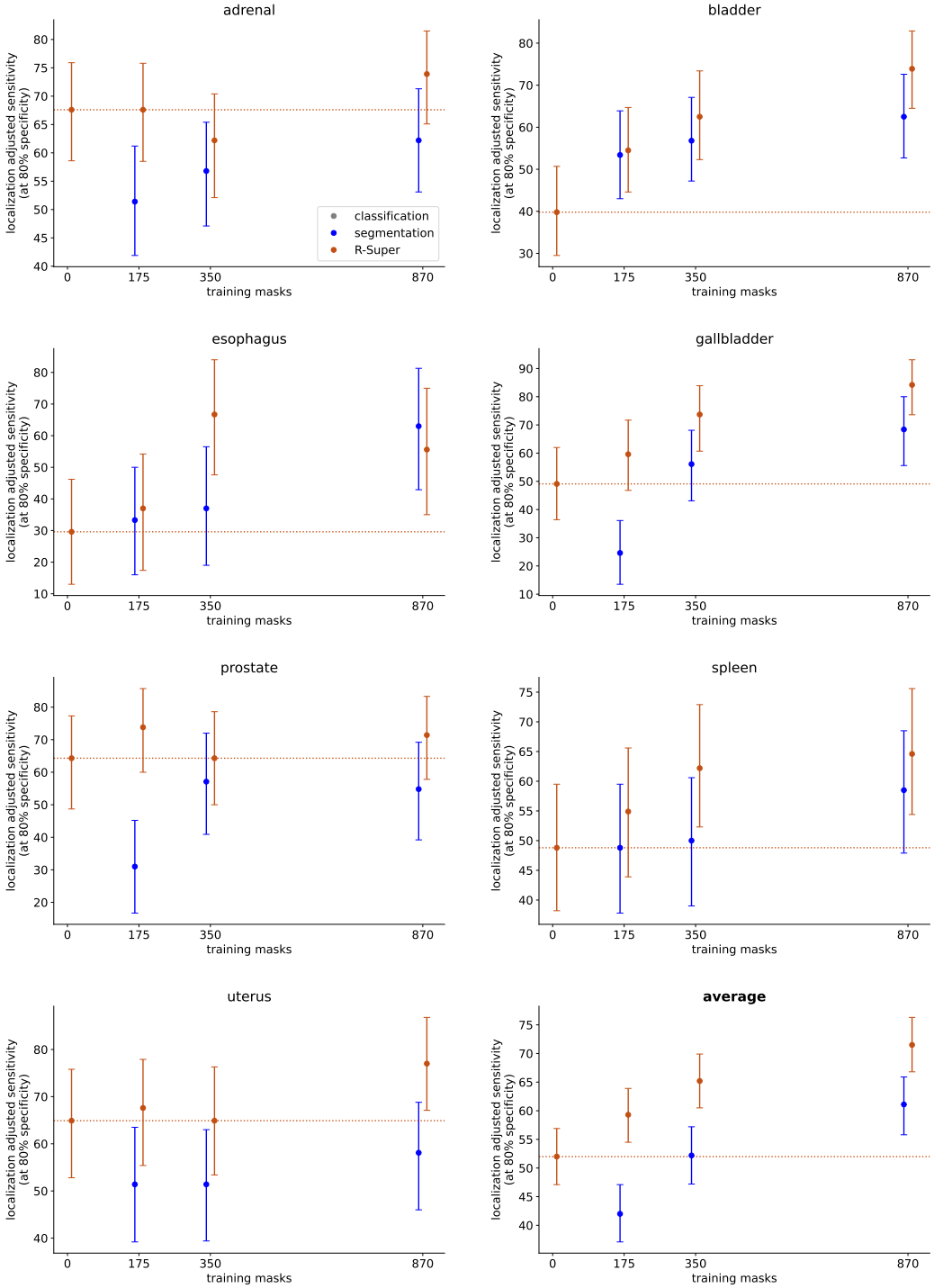

Fig. D5: **Training with many reports surpasses training with fewer masks in tumor localization, and CT-Report plus CT-Mask training consistently surpasses mask-only training.** Trained on reports alone (116,514 reports, no masks), R-Super surpassed a segmentation model trained on 175 masks and matched one trained on 350 masks in average localization adjusted sensitivity (i.e., the percentage of images where AI-made tumor segmentation masks overlap with radiologist-made tumor segmentation masks, at 80% specificity). Trained on reports and masks, R-Super† significantly surpassed the segmentation model at all mask counts. The vertical axis shows localization adjusted sensitivity for all tumors (malignant and benign); the horizontal axis indicates number of training masks. R-Super (orange) was trained on reports only (leftmost point) or reports plus masks; the segmentation model (blue) was trained on masks only. Models were trained on UCSF train and externally validated on the Stanford test set (test cases with masks, N=602). Error bars represent 95% confidence intervals, calculated with bootstrapping (1,000 bootstrap samples).

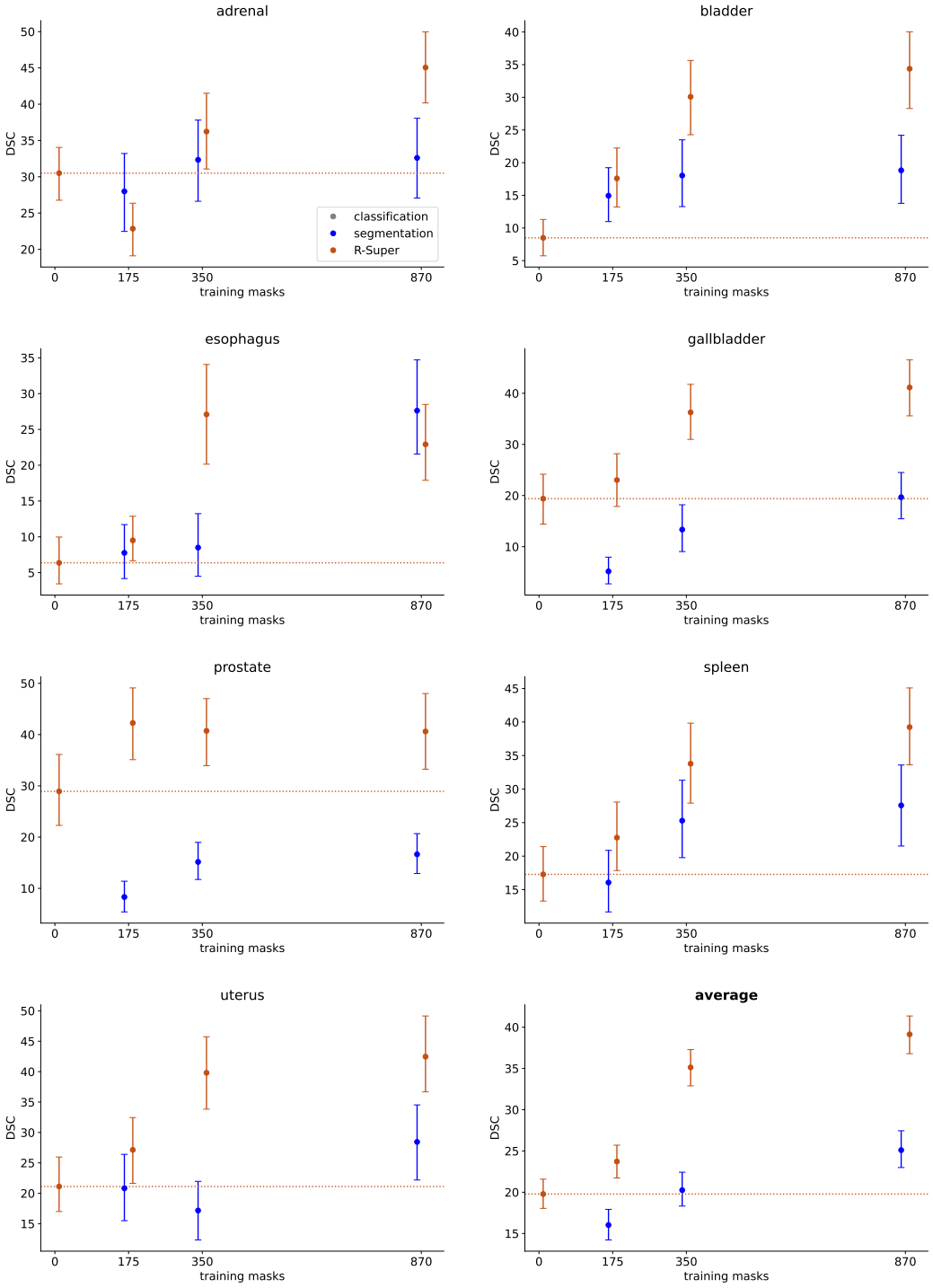

Fig. D6: **Training with many reports surpasses training with fewer masks in DSC, and CT-Report plus CT-Mask training consistently surpasses mask-only training.** The DSC values are below 50%. This is expected, since even the DSC between two radiologists reached only 56% on average (Tab. D7), given the high difficulty of detecting and segmenting these tumor types in CT. Trained on reports alone (116,514 reports, no masks), R-Super surpassed a segmentation model trained on 175 masks and matched one trained on 350 masks in average DSC. Trained on reports and masks, R-Super† significantly surpassed the segmentation model at all mask counts. The vertical axis shows DSC for tumor segmentation (malignant and benign), averaged across tumor types; the horizontal axis indicates the number of training masks. R-Super (orange) was trained on reports only (leftmost point) or reports plus masks; the segmentation model (blue) was trained on masks only. Models were trained on the UCSF train set and externally validated on the Stanford test set (test cases with masks, N=602). Error bars represent 95% confidence intervals, calculated with bootstrapping (1,000 bootstrap samples).

Table D9: **Scaling the number of training masks improves R-Super performance.** We trained multiple R-Super† models using all reports in UCSF train (117K) plus different numbers of masks. We compared to a segmentation model trained only with these masks (no report). Consistently, training with masks and reports (R-Super†) surpassed training on masks only. All models were externally evaluated on Stanford test for malignant tumor detection (malignant vs no-tumor), dataset details are in Tab. 1 and Tab. A1. All models were evaluated at 80% specificity.

training masks	bladder malignant tumor			esophagus malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
R-Super						
0	82 [71,91]	74 [64,82]	84 [76,91]	73 [50,93]	46 [28,63]	83 [68,94]
175	82 [70,91]	74 [64,82]	87 [79,93]	67 [40,90]	43 [23,60]	78 [60,93]
350	74 [62,86]	70 [59,78]	86 [80,92]	93 [79,100]	55 [36,70]	95 [90,99]
870	93 [85,98]	80 [72,87]	94 [88,98]	100 [100,100]	58 [39,71]	98 [95,100]
classification	14 [6,23]	19 [8,29]	47 [38,56]	33 [10,60]	25 [7,41]	70 [56,84]
segmentation						
175	76 [64,87]	71 [60,80]	81 [73,88]	87 [67,100]	63 [43,78]	90 [79,98]
350	66 [52,77]	64 [52,74]	78 [71,86]	73 [50,95]	46 [27,63]	82 [67,95]
870	87 [78,95]	77 [69,85]	88 [82,94]	100 [100,100]	59 [40,74]	97 [94,99]

training masks	gallbladder malignant tumor			uterus malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
R-Super						
0	83 [70,94]	70 [59,80]	92 [86,97]	88 [77,96]	78 [69,85]	95 [92,98]
175	90 [80,98]	74 [64,83]	93 [88,97]	93 [86,98]	80 [73,88]	95 [92,98]
350	85 [74,95]	71 [61,80]	91 [85,96]	77 [66,88]	72 [62,80]	90 [86,94]
870	98 [92,100]	78 [68,86]	96 [91,100]	90 [81,97]	78 [69,86]	92 [87,96]
classification	27 [14,41]	31 [17,44]	58 [46,68]	42 [28,55]	47 [33,58]	71 [62,79]
segmentation						
175	32 [18,46]	46 [28,61]	65 [58,72]	68 [57,80]	66 [56,75]	79 [71,86]
350	66 [49,80]	74 [60,84]	82 [73,89]	65 [52,77]	64 [53,74]	81 [74,88]
870	85 [72,96]	83 [74,90]	92 [85,97]	58 [44,70]	60 [48,70]	82 [75,88]

training masks	prostate malignant tumor			adrenal malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
R-Super						
0	90 [76,100]	67 [51,79]	95 [89,99]	98 [94,100]	88 [84,93]	98 [95,100]
175	90 [76,100]	61 [44,74]	96 [90,100]	98 [95,100]	88 [84,93]	98 [96,100]
350	95 [85,100]	64 [48,76]	97 [94,100]	97 [93,100]	88 [83,92]	97 [94,99]
870	95 [84,100]	64 [48,75]	97 [93,100]	98 [95,100]	89 [84,94]	98 [95,99]
classification	67 [46,86]	49 [32,64]	79 [68,88]	22 [14,31]	31 [21,40]	55 [47,62]
segmentation						
175	48 [27,68]	38 [21,53]	69 [57,80]	90 [84,96]	93 [89,96]	95 [91,98]
350	71 [50,90]	52 [35,66]	87 [79,93]	95 [90,99]	94 [90,98]	97 [94,99]
870	86 [70,100]	60 [44,74]	88 [78,97]	92 [86,97]	93 [89,96]	96 [93,98]

training masks	spleen malignant tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
R-Super						
0	90 [83,96]	82 [74,88]	92 [88,96]	86 [82,91]	72 [67,77]	91 [89,94]
175	97 [93,100]	85 [79,90]	96 [93,98]	88 [83,92]	72 [67,77]	92 [89,94]
350	94 [89,99]	83 [76,89]	95 [92,98]	88 [84,92]	72 [67,76]	93 [92,95]
870	97 [93,100]	85 [79,91]	96 [93,99]	96 [93,98]	76 [71,80]	96 [94,97]
classification	25 [14,35]	32 [20,42]	55 [46,64]	33 [27,39]	33 [28,38]	62 [58,66]
segmentation						
175	81 [72,89]	88 [82,93]	90 [86,95]	69 [64,74]	66 [61,71]	81 [78,84]
350	80 [70,89]	76 [68,83]	87 [81,92]	74 [68,79]	67 [62,72]	85 [82,88]
870	87 [78,94]	86 [79,91]	92 [87,96]	85 [81,89]	74 [69,78]	91 [88,93]

Table D10: **Scaling the number of training reports improves R-Super performance.** We trained multiple R-Super† models using all masks in UCSF train plus different numbers of reports. We also trained multiple R-Super models using no mask and different numbers of reports. In both cases, increasing the number of training reports improved performance. All models were externally evaluated on Stanford test for malignant tumor detection (malignant vs no-tumor), dataset details are in Tab. 1 and Tab. A1. All models were evaluated at 80% specificity.

training reports	bladder malignant tumor			esophagus malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
R-Super (reports & all masks)						
1k reports	56 [42,70]	57 [45,68]	78 [70,84]	93 [78,100]	55 [35,71]	95 [90,99]
10k reports	78 [68,89]	72 [62,81]	90 [84,95]	100 [100,100]	58 [40,72]	97 [93,99]
all reports (116k)	93 [85,98]	80 [72,87]	94 [88,98]	100 [100,100]	58 [39,71]	98 [95,100]
R-Super (reports & no mask)						
1k reports	62 [48,75]	61 [50,72]	75 [66,82]	33 [10,57]	24 [7,41]	66 [51,80]
10k reports	29 [17,41]	34 [21,46]	70 [61,77]	80 [58,100]	50 [30,66]	90 [80,98]
all reports (116k)	82 [71,91]	74 [64,82]	84 [76,91]	73 [50,93]	46 [28,63]	83 [68,94]

training reports	gallbladder malignant tumor			uterus malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
R-Super (reports & all masks)						
1k reports	80 [68,92]	69 [57,79]	85 [77,92]	74 [62,85]	69 [61,79]	86 [81,91]
10k reports	93 [84,100]	75 [65,84]	94 [89,98]	61 [49,74]	61 [50,71]	87 [82,92]
all reports (116k)	98 [92,100]	78 [68,86]	96 [91,100]	90 [81,97]	78 [69,86]	92 [87,96]
R-Super (reports & no mask)						
1k reports	37 [22,52]	40 [25,54]	63 [52,73]	72 [59,84]	68 [57,77]	87 [82,92]
10k reports	90 [80,98]	74 [64,83]	94 [88,98]	65 [52,77]	64 [53,74]	87 [81,92]
all reports (116k)	83 [70,94]	70 [59,80]	92 [86,97]	88 [77,96]	78 [69,85]	95 [92,98]

training reports	prostate malignant tumor			adrenal malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
R-Super (reports & all masks)						
1k reports	86 [70,100]	59 [43,73]	93 [88,97]	97 [93,100]	88 [83,92]	98 [95,99]
10k reports	90 [77,100]	61 [45,74]	92 [85,97]	98 [95,100]	88 [84,93]	99 [97,100]
all reports (116k)	95 [84,100]	64 [48,75]	97 [93,100]	98 [95,100]	89 [84,94]	98 [95,99]
R-Super (reports & no mask)						
1k reports	86 [68,100]	59 [43,72]	92 [86,96]	42 [32,52]	50 [41,59]	65 [58,72]
10k reports	81 [62,95]	57 [40,71]	92 [86,97]	76 [67,85]	76 [69,82]	86 [80,91]
all reports (116k)	90 [76,100]	67 [51,79]	95 [89,99]	98 [94,100]	88 [84,93]	98 [95,100]

training reports	spleen malignant tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
R-Super (reports & all masks)						
1k reports	94 [88,99]	83 [76,89]	94 [90,98]	83 [79,87]	69 [64,74]	90 [88,92]
10k reports	94 [88,99]	83 [76,89]	95 [91,98]	88 [85,91]	71 [66,76]	93 [92,95]
all reports (116k)	97 [93,100]	85 [79,91]	96 [93,99]	96 [93,98]	76 [71,80]	96 [94,97]
R-Super (reports & no mask)						
1k reports	30 [19,42]	38 [26,50]	56 [47,65]	52 [46,57]	49 [43,53]	72 [68,75]
10k reports	30 [20,42]	38 [26,49]	57 [48,66]	64 [60,69]	56 [51,61]	82 [79,85]
all reports (116k)	90 [83,96]	82 [74,88]	92 [88,96]	86 [82,91]	72 [67,77]	91 [89,94]

D.4 LLM Evaluation

Table D11: **The LLM achieves high accuracy in extracting tumor information from reports.** Radiologists evaluated the LLM outputs on 448 radiology reports, of which 130 had paired pathology reports. *The tumor detection and localization evaluation* assess whether each LLM-extracted tumor is truly mentioned in the report, and is assigned to the correct organ; it also assesses whether reports the LLM deemed tumor-free across the 7 organs were truly tumor-free. *The malignancy, size, and attenuation evaluation* assess whether each LLM-extracted tumor malignancy/size/attenuation matched the report. Evaluation cohorts: tumor detection and localization used the full dataset (tumor reports: 47 adrenal; 48 bladder; 27 esophagus; 41 gallbladder; 57 prostate; 28 spleen; 56 uterus; 191 no-tumor reports); malignancy was evaluated for all correctly extracted tumors; size for 219 tumors (others had no reported size) and attenuation for 92 tumors (others had no reported attenuation). Sensitivity, specificity, and F1-score are reported only for tumor detection and localization, because the other tasks have no negative class (each lesion yields a single right/wrong verdict, unlike detection, which has a no-tumor class). Results are shown as average and 95% CI (bootstrap, 1,000 samples).

metric	adrenal	bladder	esophagus	gallbladder
tumor detection and localization				
accuracy	99.5 [98.6,100.0]	99.5 [98.6,100.0]	99.0 [97.5,100.0]	98.6 [96.7,100.0]
sensitivity	97.9 [92.9,100.0]	97.9 [92.9,100.0]	100.0 [100.0,100.0]	95.1 [87.8,100.0]
specificity	100.0 [100.0,100.0]	100.0 [100.0,100.0]	98.8 [97.0,100.0]	99.4 [98.2,100.0]
F1	98.9 [96.3,100.0]	98.9 [96.3,100.0]	96.4 [90.0,100.0]	96.3 [91.7,100.0]
malignancy				
accuracy	93.3 [84.4,100.0]	100.0 [100.0,100.0]	96.3 [88.9,100.0]	92.5 [85.0,100.0]
size				
accuracy	97.5 [92.5,100.0]	100.0 [100.0,100.0]	100.0 [100.0,100.0]	89.7 [79.3,100.0]
attenuation				
accuracy	90.9 [72.7,100.0]	90.0 [70.0,100.0]	83.3 [50.0,100.0]	93.3 [80.0,100.0]

metric	prostate	spleen	uterus	average
tumor detection and localization				
accuracy	94.9 [91.2,97.7]	98.5 [96.5,100.0]	98.2 [96.4,99.6]	98.3 [97.6,98.9]
sensitivity	80.7 [69.2,89.8]	96.4 [87.1,100.0]	92.9 [85.7,98.3]	94.4 [91.9,96.6]
specificity	100.0 [100.0,100.0]	98.8 [97.1,100.0]	100.0 [100.0,100.0]	99.6 [99.2,99.9]
F1	89.3 [81.8,94.6]	94.7 [87.5,100.0]	96.3 [92.3,99.2]	95.8 [94.0,97.4]
malignancy				
accuracy	91.3 [82.6,97.8]	85.2 [70.4,96.3]	100.0 [100.0,100.0]	94.1 [91.2,96.7]
size				
accuracy	89.5 [78.9,97.4]	100.0 [100.0,100.0]	92.5 [82.5,100.0]	95.6 [93.1,97.9]
attenuation				
accuracy	92.9 [78.6,100.0]	100.0 [100.0,100.0]	100.0 [100.0,100.0]	92.9 [86.3,98.7]

We evaluated the LLM's performance at extracting tumor information from reports. The evaluation covered four tasks: *(i)* tumor detection and localization, i.e., determining whether each LLM-extracted tumor truly existed in the report and was assigned to the correct organ, and whether reports deemed tumor-free across the seven organs were truly tumor-free; *(ii)* malignancy extraction, i.e., determining whether the LLM-extracted tumor malignancy matched the report; *(iii)* size extraction, i.e., determining whether the LLM-extracted tumor diameters matched the report; and *(iv)* attenuation extraction, i.e., determining whether the LLM-extracted attenuation matched the report. Results are shown in Tab. D11. We evaluated the LLM on 448 radiology reports, including 130 cases with paired pathology reports (used for malignancy extraction). To assess generalization, the evaluation included reports from multiple hospitals: 67% came from UCSF-test (including all pathology reports), 18% from Medipol, and 15% from Stanford-Test.

Our LLM prompt (Supp. G) and this evaluation use a broad definition of tumor, which is in fact closer to a clinical definition of lesion: "any focal abnormality, including masses, cysts, or areas of altered density (hyperdense, hypodense, or isodense)." This definition is intentionally broader than the standard clinical tumor definition because it is designed to capture any lesion that may be concerning for current or future malignancy. Accordingly, the definition includes benign tumors, metastases, primary malignant tumors, nodules, soft tissue lesions, tissue thickening, lesions, masses, cysts, pseudo-cysts, neoplasms, cancers, index lesions, oncologic findings, adenomas, carcinomas, growths, abnormalities, foci, LI-RADS lesions, hyperdensities, hypodensities, and isodensities. The LLM is instructed to extract both certain and uncertain lesions, recording both the tumor type and the certainty according to the reports.

For tumor detection and localization, the LLM achieved 98.3% average accuracy (95% CI: 97.6-98.9) and 95.8% average F1-Score (94.0-97.4). This performance surpasses natural language processing algorithms used to process reports and extract ground-truth labels in widely used datasets such as ChestX-ray8 (90% F1-Score [50]). In our evaluation, an LLM-extracted tumor was considered a true positive if the tumor was described in the report and the LLM assigned it to the correct organ. An extracted tumor not described in the report for that organ was considered a false positive. Most false positives were not random, but borderline findings. Examples included fluid collections adjacent to the spleen being classified as spleen lesions and distended gallbladders without reported lesions being classified as lesions. To identify possible false negatives, radiologists also reviewed

219 reports for which the LLM extracted no tumor in any of the seven organs. False negatives also involved borderline findings, such as minor spleen calcifications being ignored by the LLM. Importantly, none of the false negatives were malignant tumors, meaning that the LLM successfully captured all malignant tumors in this evaluation dataset.

For malignancy extraction, the LLM achieved 94.1% accuracy (91.2-96.7%). Our broad tumor definition is appropriate because R-Super not only detects tumors but also estimates whether they are malignant, which is a much narrower definition. The LLM extracts malignancy information from pathology reports when available and from radiology reports otherwise. The LLM also determines whether malignancy is certain (e.g., a clear pathology result or a mention of known active cancer diagnosis in a radiology report) or uncertain. In this evaluation dataset, 41% of lesions were malignant, 36% were benign, and 23% were uncertain. Among the 130 cases with pathology reports, only a single LLM error was identified. In that case, the pathology report confirmed a right adrenal metastasis, and the LLM incorrectly classified an additional right adrenal nodule as malignant, even though the radiology report described it as indeterminate.

For tumor size extraction, the LLM achieved 95.6% accuracy (93.1-98.0%). A total of 219 tumors had reported sizes. Most LLM errors occurred when the LLM extracted the organ size instead of the tumor size. For example, one report stated, "The uterus is enlarged, measuring 14.2 × 8.3 × 8.6 cm with a large mass centered within the uterine body and fundus" and the LLM incorrectly extracted 14.2 × 8.3 × 8.6 cm as the tumor diameter.

For tumor attenuation extraction, the LLM achieved 92.9% accuracy (86.2-98.6%). A total of 92 tumors had reported attenuation. Errors included inferring attenuation directly from tumor type (e.g., classifying an adrenal adenoma as low attenuation) and misinterpreting PET hypermetabolism as CT enhancement.

Overall, the LLM achieved high performance (Tab. D11), surpassing widely accepted natural language processing algorithms used to extract information from reports and label medical datasets [50, 56]. The LLM's strong performance across reports from three hospitals in two countries (USA and Turkey) further demonstrates robust generalization. Nevertheless, the LLM accuracy was not perfect, indicating that some LLM errors exist in the R-Super training pipeline. Despite these errors, R-Super outperformed six alternative training methods (not LLM-dependent) and multiple public AI models, demonstrating that R-Super is robust to occasional LLM extraction errors.

D.5 Ablation Studies

We perform ablation studies to demonstrate that each of our loss functions contributes to model performance. Specifically, we trained versions of R-Super† without the Ball Loss, Volume Loss, or Attenuation Loss. We did not ablate the Malignancy Loss because its purpose is to enable R-Super to distinguish between malignant and benign tumors, rather than improve tumor detection. I.e., if we remove the Malignancy Loss, R-Super simply loses its ability to identify malignancy. In the full R-Super, the Volume Loss is applied as deep supervision and the Ball Loss at the final layer. In the ablations, when either of these losses was removed, the remaining loss was applied both as deep supervision and at the final layer.

The ablation results (Tab. D13 to D12) show that the Ball Loss, Volume Loss, and Attenuation Loss all contribute to tumor detection performance. Removing any of these losses reduces average tumor detection AUC and F1-score across the three external test sets. Only in the internal UCSF test set, removing the Attenuation Loss did not affect the average performance. Consistently, removing the Attenuation Loss produces the smallest AUC decrease, whereas removing the Volume Loss produces the largest decrease.

These findings are consistent with the design of the Ball Loss and Volume Loss. The Ball Loss refines the segmentation model's output according to the tumor description in the radiology report, and the refined output guides the model updates. Therefore, the Ball Loss depends on the current segmentation model's output. Early in training this output is still largely random, and the resulting loss signal is noisy and difficult to optimize. In contrast, the Volume Loss does not attempt to refine the segmentation model's output. Instead, it enforces that segmented tumors have the volume described in the report and are located in the correct organs. Consequently, the Volume Loss provides a more stable training signal when model outputs are still unreliable, leading to better convergence. This explains why training with the Volume Loss alone surpasses training with the Ball Loss alone. The best performance is achieved when both losses are used together: the Volume Loss improves training convergence, especially when the segmentation model's outputs are unreliable; once these outputs become reliable, the Ball Loss further improves performance by ensuring that the segmented tumors match the report not only in volume and location, but also in number and diameter.

Table D12: **Ablation studies (internal validation, UCSF, USA): removing the Volume Loss or the Ball Loss decreased performance.** Unlike the 3 external test sets, removing the Attenuation Loss did not reduce average tumor detection AUC and F1-Score in internal validation. This table presents ablation studies of R-Super† for tumor detection (tumor vs no-tumor). All models were trained with CT-Report and CT-Mask pairs from the R-Super Train dataset (Tab. A1). We compare the full R-Super to versions trained without the Ball Loss, Volume Loss, or Attenuation Loss. Removing the Volume Loss causes the largest AUC decrease, followed by the Ball Loss removal. Train/test dataset details are provided in Tab. A1. Results are shown as average and 95% confidence intervals, calculated via bootstrapping with 1,000 bootstrap samples. Blue indicates statistical significance ($p < 0.05$). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics.

AI	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	52 [47,57]	92 [87,97]	67 [62,72]	76 [72,81]	82 [78,86]	94 [89,98]	89 [87,92]	94 [91,95]
w/o ball loss	66 [61,71]	86 [79,92]	78 [74,81]	82 [78,86]	82 [77,85]	94 [89,98]	89 [86,91]	94 [92,96]
w/o att. loss	77 [72,81]	67 [58,76]	82 [78,85]	78 [73,82]	82 [78,85]	95 [90,98]	89 [86,91]	94 [91,96]
R-Super†	64 [59,69]	91 [85,96]	76 [73,80]	82 [78,86]	82 [79,86]	90 [85,95]	89 [86,91]	90 [88,93]

AI	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	66 [60,72]	77 [69,84]	75 [71,79]	77 [72,81]	83 [78,87]	85 [78,91]	88 [85,91]	87 [84,91]
w/o ball loss	70 [65,76]	76 [68,84]	78 [74,82]	78 [73,82]	78 [73,83]	86 [79,92]	85 [82,88]	86 [82,89]
w/o att. loss	63 [57,68]	89 [83,94]	75 [70,79]	82 [77,86]	85 [81,89]	88 [81,93]	90 [87,92]	88 [85,92]
R-Super†	78 [73,83]	80 [73,87]	83 [80,87]	84 [80,88]	80 [75,84]	90 [85,95]	87 [84,90]	88 [84,91]

AI	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	82 [77,88]	72 [63,80]	82 [77,86]	82 [77,87]	81 [78,84]	96 [92,99]	89 [87,91]	90 [88,92]
w/o ball loss	83 [78,88]	76 [68,84]	84 [79,88]	84 [80,89]	85 [83,88]	88 [83,94]	91 [89,93]	92 [90,94]
w/o att. loss	88 [83,92]	75 [66,83]	86 [82,90]	84 [78,88]	86 [83,88]	94 [89,98]	92 [90,93]	92 [91,94]
R-Super†	86 [81,90]	80 [71,87]	86 [82,90]	87 [82,91]	82 [79,85]	92 [87,97]	90 [88,91]	91 [89,93]

AI	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	66 [62,71]	96 [92,99]	80 [76,83]	85 [82,88]	73 [71,75]	87 [85,89]	82 [80,83]	84 [83,86]
w/o ball loss	64 [60,68]	90 [85,96]	77 [74,80]	80 [76,84]	76 [74,77]	85 [83,88]	83 [82,84]	85 [84,87]
w/o att. loss	72 [68,76]	96 [91,99]	83 [80,86]	89 [86,91]	79 [77,80]	86 [84,89]	85 [84,86]	87 [85,88]
R-Super†	75 [71,79]	95 [90,98]	85 [82,88]	87 [84,90]	78 [76,80]	88 [86,90]	85 [84,86]	87 [86,88]

Table D13: **Ablation studies (external validation, Stanford, USA): all three loss functions contribute to performance.** This table presents ablation studies of R-Super† for tumor detection (tumor vs no-tumor). All models were trained with the CT-Report and CT-Mask pairs from the UCSF Train dataset. We compare the full R-Super to versions trained without the Ball Loss, Volume Loss, or Attenuation Loss. Removing any loss function reduces tumor detection performance, demonstrating that all three losses contribute to the final model. Consistent with the design of the method, removing the Volume Loss causes the largest AUC decrease, whereas removing the Attenuation Loss causes the smallest decrease. Train/test dataset details are provided in Tab. A1. Results are shown as average and 95% confidence intervals, calculated via bootstrapping with 1,000 bootstrap samples. Blue indicates statistical significance ($p < 0.05$). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics.

AI	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	68 [61,74]	87 [80,92]	77 [72,82]	81 [76,86]	68 [53,82]	95 [90,99]	75 [62,85]	85 [76,92]
w/o ball loss	78 [72,83]	76 [68,84]	81 [77,86]	82 [77,86]	85 [74,95]	84 [77,90]	74 [63,84]	91 [85,96]
w/o att. loss	68 [61,74]	90 [85,95]	78 [73,83]	84 [80,89]	83 [71,93]	80 [72,86]	69 [57,79]	85 [77,92]
R-Super†	72 [66,79]	84 [77,90]	80 [75,84]	84 [79,88]	78 [65,91]	92 [87,96]	78 [67,87]	90 [83,95]

AI	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	64 [57,70]	90 [83,95]	75 [70,80]	81 [76,85]	89 [85,92]	67 [58,75]	89 [86,91]	85 [81,89]
w/o ball loss	64 [57,70]	91 [86,96]	76 [70,81]	81 [76,85]	81 [77,85]	71 [63,79]	85 [82,88]	84 [79,88]
w/o att. loss	67 [60,74]	88 [82,94]	77 [72,82]	84 [79,88]	77 [73,81]	79 [71,86]	84 [81,86]	86 [82,90]
R-Super†	76 [70,81]	89 [83,94]	83 [79,87]	87 [83,91]	84 [80,87]	79 [71,86]	88 [85,90]	89 [84,92]

AI	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	81 [73,88]	75 [67,82]	78 [72,84]	88 [84,92]	78 [76,81]	91 [86,96]	87 [85,89]	90 [87,92]
w/o ball loss	73 [64,81]	90 [83,94]	79 [73,85]	90 [87,94]	78 [75,81]	83 [76,90]	86 [84,88]	88 [85,91]
w/o att. loss	73 [65,82]	88 [81,93]	79 [72,84]	90 [86,94]	76 [74,79]	97 [94,100]	86 [84,88]	92 [90,94]
R-Super†	94 [89,98]	68 [60,76]	82 [77,88]	90 [86,93]	83 [80,85]	90 [84,95]	90 [88,91]	90 [88,93]

AI	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	72 [68,76]	96 [92,99]	83 [80,86]	82 [78,85]	74 [71,77]	86 [84,88]	81 [78,83]	84 [83,86]
w/o ball loss	52 [48,56]	96 [92,99]	68 [64,72]	79 [75,83]	73 [70,75]	85 [82,87]	79 [76,81]	85 [83,86]
w/o att. loss	77 [73,80]	91 [86,96]	86 [84,88]	86 [83,89]	74 [72,77]	88 [85,90]	80 [78,82]	87 [85,89]
R-Super†	76 [73,80]	90 [83,95]	85 [83,88]	83 [80,86]	80 [78,83]	84 [82,87]	84 [82,86]	88 [86,89]

Table D14: **Ablation studies (external validation, Medipol, Turkey): all three loss functions contribute to performance.** This table presents ablation studies of R-Super† for tumor detection (tumor vs no-tumor). All models were trained with the CT-Report and CT-Mask pairs from the R-Super Train dataset. We compare the full R-Super to versions trained without the Ball Loss, Volume Loss, or Attenuation Loss. Removing any loss function reduces tumor detection performance, demonstrating that all three losses contribute to the final model. Consistent with the design of the method, removing the Volume Loss causes the largest AUC decrease, whereas removing the Attenuation Loss causes the smallest decrease. Train/test dataset details are provided in Tab. A1. Results are shown as average and 95% confidence intervals, calculated via bootstrapping with 1,000 bootstrap samples. Blue indicates statistical significance ($p < 0.05$). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics.

AI	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	77 [71,84]	72 [65,79]	73 [67,79]	82 [77,86]	52 [45,60]	96 [92,98]	67 [60,73]	80 [75,85]
w/o ball loss	81 [74,87]	82 [76,87]	80 [74,84]	87 [83,90]	58 [51,66]	92 [88,96]	70 [64,76]	79 [75,84]
w/o att. loss	80 [73,86]	69 [62,75]	74 [68,78]	80 [75,85]	59 [51,66]	87 [82,92]	69 [62,74]	82 [78,86]
R-Super†	61 [53,69]	94 [90,97]	72 [66,78]	87 [83,90]	83 [78,88]	80 [74,86]	82 [78,86]	89 [86,92]

AI	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	36 [29,44]	90 [85,94]	49 [42,57]	64 [58,69]	94 [90,98]	66 [59,73]	78 [73,83]	89 [86,93]
w/o ball loss	45 [38,52]	92 [88,96]	59 [52,65]	72 [66,77]	92 [88,96]	72 [66,79]	80 [75,85]	90 [87,93]
w/o att. loss	60 [53,68]	80 [74,85]	67 [61,72]	73 [67,78]	79 [72,86]	86 [81,91]	80 [74,85]	91 [87,94]
R-Super†	67 [60,74]	84 [78,89]	73 [68,78]	77 [72,82]	95 [90,98]	68 [62,76]	80 [75,84]	90 [87,93]

AI	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	100 [100,100]	79 [73,84]	65 [54,75]	96 [93,98]	74 [69,78]	92 [88,96]	83 [80,86]	85 [82,88]
w/o ball loss	94 [85,100]	93 [89,96]	82 [71,90]	96 [94,99]	81 [77,85]	91 [87,95]	88 [85,90]	88 [85,91]
w/o att. loss	86 [73,96]	97 [94,99]	84 [74,93]	95 [91,98]	78 [74,82]	94 [91,98]	86 [84,89]	91 [88,94]
R-Super†	89 [77,98]	97 [94,99]	86 [76,94]	96 [92,99]	82 [78,86]	92 [87,96]	88 [86,91]	90 [88,93]

AI	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	75 [71,79]	75 [69,81]	80 [77,83]	81 [77,85]	73 [71,75]	81 [79,84]	71 [68,73]	82 [81,84]
w/o ball loss	74 [69,78]	71 [64,78]	78 [75,82]	76 [72,81]	75 [73,77]	85 [83,87]	77 [74,79]	84 [83,86]
w/o att. loss	71 [66,76]	83 [77,88]	79 [75,82]	81 [77,85]	73 [71,76]	85 [83,87]	77 [74,79]	85 [83,86]
R-Super†	69 [65,74]	82 [76,88]	78 [74,81]	80 [76,84]	78 [75,81]	85 [83,88]	80 [78,82]	87 [86,89]

Table D15: **Ablation studies (external validation, Basel, Switzerland): all three loss functions contribute to performance.** This table presents ablation studies of R-Super† for tumor detection (tumor vs no-tumor). All models were trained with the CT-Report and CT-Mask pairs from the R-Super Train dataset. We compare the full R-Super to versions trained without the Ball Loss, Volume Loss, or Attenuation Loss. Removing any loss function reduces tumor detection performance, demonstrating that all three losses contribute to the final model. Consistent with the design of the method, removing the Volume Loss causes the largest AUC decrease, whereas removing the Attenuation Loss causes the smallest decrease. Train/test dataset details are provided in Tab. A1. Results are shown as average and 95% confidence intervals, calculated via bootstrapping with 1,000 bootstrap samples. Blue indicates statistical significance ($p < 0.05$). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics.

AI	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	80 [76,84]	95 [88,100]	88 [86,91]	92 [89,95]	86 [78,93]	91 [83,98]	90 [84,94]	90 [84,94]
w/o ball loss	81 [77,86]	96 [91,100]	89 [86,92]	92 [89,95]	84 [75,90]	86 [76,94]	87 [81,92]	91 [86,95]
w/o att. loss	78 [73,83]	91 [83,98]	86 [83,90]	91 [86,94]	86 [79,93]	91 [83,98]	90 [85,94]	93 [88,97]
R-Super†	84 [80,89]	93 [85,98]	91 [88,93]	92 [88,94]	90 [83,96]	86 [77,94]	90 [85,94]	92 [87,96]

AI	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	37 [26,48]	93 [85,98]	52 [40,62]	67 [57,76]	84 [80,89]	98 [94,100]	91 [88,94]	91 [88,94]
w/o ball loss	71 [60,81]	75 [64,87]	75 [66,82]	76 [68,84]	84 [79,88]	91 [83,98]	90 [87,93]	91 [87,94]
w/o att. loss	63 [52,74]	89 [81,97]	74 [64,82]	80 [72,87]	84 [79,88]	95 [88,100]	90 [87,93]	90 [86,93]
R-Super†	55 [43,67]	98 [94,100]	71 [60,80]	83 [76,89]	81 [76,86]	96 [91,100]	89 [86,92]	90 [86,93]

AI	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	88 [83,93]	79 [67,88]	90 [86,93]	86 [80,91]	76 [73,78]	95 [88,100]	86 [84,87]	87 [83,90]
w/o ball loss	82 [76,88]	79 [67,89]	87 [82,90]	88 [83,92]	76 [73,78]	93 [85,98]	86 [84,87]	87 [83,91]
w/o att. loss	86 [80,91]	82 [72,91]	89 [85,92]	84 [78,90]	78 [76,80]	86 [76,95]	88 [86,89]	87 [84,91]
R-Super†	87 [82,92]	79 [67,88]	90 [86,93]	87 [81,91]	77 [75,79]	93 [85,98]	87 [86,88]	87 [84,90]

AI	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
w/o volume loss	59 [55,62]	89 [81,96]	74 [70,76]	78 [73,83]	73 [70,75]	91 [88,94]	82 [79,84]	84 [82,86]
w/o ball loss	62 [59,66]	61 [48,73]	75 [72,78]	60 [52,68]	77 [75,79]	83 [80,86]	84 [82,86]	84 [82,85]
w/o att. loss	63 [59,66]	75 [63,86]	76 [73,79]	72 [64,78]	77 [74,79]	87 [84,90]	85 [83,86]	85 [83,87]
R-Super†	77 [74,80]	80 [70,90]	86 [84,88]	82 [77,86]	79 [76,81]	89 [86,92]	86 [84,88]	87 [86,89]

D.6 Comparison to Public AI Models

In this Section, we compare R-Super, trained only with CT-Report pairs, to state-of-the-art public AI models, including segmentation models and foundational vision-language model. We demonstrate that these public models have difficulty detecting the 7 tumor types studied here, which are absent from previous public segmentation datasets. Consistently, R-Super surpasses all previous public AI models in tumor detection, malignant tumor detection, and small tumor detection.

D.6.1 Tumor Detection

Table D16: **Tumor detection, internal validation (UCSF, USA): trained with no mask, R-Super surpasses state-of-the-art public AI models in the detection of 7 malignant tumor types.** This table evaluates tumor detection (malignant + benign vs no-tumor). R-Super was trained with the R-Super train set (USA, no tumor mask), and it was evaluated on the UCSF test set (internal validation). See dataset details in Tab. 1, A1. We compare R-Super to prominent public AI models: MedGemma, Merlin, MedImageInsight, Lingshu, ULS, and winners of the FLARE cancer segmentation challenge. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model for each metric. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples.

AI	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	70 [66,75]	41 [32,50]	74 [70,78]	55 [50,62]	45 [40,50]	80 [73,88]	60 [55,64]	64 [58,70]
Merlin	43 [37,48]	81 [74,88]	58 [52,63]	64 [59,70]	62 [58,67]	69 [60,77]	73 [69,77]	70 [65,75]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,51]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	49 [48,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	28 [23,33]	79 [71,86]	41 [35,47]	52 [47,58]	4 [3,7]	98 [95,100]	8 [5,12]	50 [46,53]
FLARE23 Top 1	4 [2,6]	99 [97,100]	8 [5,12]	51 [50,53]	0 [0,1]	100 [100,100]	1 [0,3]	50 [50,51]
FLARE23 Top 2	4 [2,5]	99 [97,100]	7 [3,10]	51 [50,52]	10 [8,13]	99 [97,100]	19 [14,23]	55 [53,56]
FLARE24 Top 1	7 [4,10]	99 [97,100]	13 [9,18]	53 [51,55]	3 [1,4]	100 [100,100]	6 [3,9]	51 [51,52]
FLARE25 Baseline	10 [7,14]	97 [94,100]	18 [13,24]	54 [52,56]	5 [3,7]	100 [100,100]	9 [5,14]	52 [51,54]
segmentation from reports (ours)								
R-Super	77 [72,81]	73 [64,80]	83 [79,86]	79 [74,83]	85 [81,88]	87 [80,93]	90 [87,92]	91 [88,94]

AI	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	66 [60,71]	44 [35,53]	70 [65,74]	56 [49,62]	51 [46,56]	57 [48,65]	61 [56,66]	51 [46,57]
Merlin	29 [23,35]	80 [73,88]	42 [36,48]	52 [46,58]	56 [50,62]	77 [69,84]	68 [63,73]	68 [63,73]
VLM report generation								
Merlin	1 [0,2]	100 [100,100]	2 [0,4]	50 [50,51]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	2 [0,3]	100 [100,100]	3 [1,6]	51 [50,52]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	50 [47,52]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	8 [5,11]	96 [92,99]	15 [9,20]	51 [49,54]	25 [20,30]	85 [78,91]	38 [32,44]	54 [50,59]
FLARE23 Top 1	18 [13,22]	98 [95,100]	30 [23,36]	58 [55,60]	6 [3,8]	100 [100,100]	11 [6,15]	53 [52,54]
FLARE23 Top 2	43 [37,48]	95 [90,98]	59 [53,65]	69 [65,72]	7 [5,11]	100 [100,100]	14 [9,19]	54 [52,55]
FLARE24 Top 1	22 [18,27]	98 [95,100]	36 [30,42]	60 [57,63]	23 [18,28]	100 [100,100]	38 [31,44]	62 [59,64]
FLARE25 Baseline	20 [16,25]	98 [95,100]	33 [27,40]	59 [56,62]	40 [34,45]	92 [87,97]	56 [50,61]	66 [62,69]
segmentation from reports (ours)								
R-Super	73 [67,78]	72 [63,79]	79 [74,82]	78 [73,82]	83 [78,87]	87 [80,93]	88 [85,91]	89 [86,92]

AI	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	63 [56,70]	54 [46,63]	66 [60,72]	57 [51,64]	66 [63,70]	45 [36,54]	76 [73,78]	54 [48,59]
Merlin	64 [57,71]	74 [65,81]	71 [65,76]	70 [64,76]	42 [38,46]	78 [70,85]	58 [54,62]	64 [58,69]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	3 [2,4]	98 [96,100]	6 [4,8]	51 [49,52]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,1]	50 [50,50]
segmentation from masks								
ULS	26 [19,32]	90 [84,96]	39 [31,47]	58 [53,63]	1 [0,2]	100 [100,100]	2 [1,3]	50 [50,51]
FLARE23 Top 1	1 [0,2]	100 [100,100]	1 [0,4]	50 [50,51]	11 [9,14]	100 [100,100]	20 [16,24]	56 [54,57]
FLARE23 Top 2	1 [0,2]	100 [100,100]	1 [0,3]	50 [49,51]	27 [23,30]	100 [100,100]	42 [38,46]	63 [62,65]
FLARE24 Top 1	2 [0,4]	100 [100,100]	3 [0,7]	50 [48,52]	11 [8,13]	100 [100,100]	19 [15,23]	55 [54,56]
FLARE25 Baseline	24 [18,30]	93 [88,97]	37 [29,44]	58 [54,62]	10 [8,12]	100 [100,100]	18 [15,22]	55 [54,56]
segmentation from reports (ours)								
R-Super	69 [63,76]	92 [87,97]	80 [74,84]	85 [80,90]	87 [85,90]	88 [81,93]	92 [90,94]	93 [90,95]

AI	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	71 [67,75]	42 [33,50]	77 [73,80]	56 [50,62]	62 [60,64]	52 [47,57]	69 [67,71]	56 [53,60]
Merlin	58 [53,62]	58 [48,67]	69 [65,72]	59 [54,65]	51 [48,53]	74 [68,80]	63 [60,65]	64 [60,68]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,51]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	10 [7,12]	95 [90,98]	17 [13,22]	52 [49,55]	2 [2,2]	99 [98,100]	4 [3,5]	50 [50,51]
Lingshu	27 [23,31]	77 [69,84]	41 [36,46]	52 [48,56]	4 [3,4]	97 [96,98]	6 [5,7]	50 [49,51]
segmentation from masks								
ULS	22 [18,26]	89 [84,95]	36 [30,41]	56 [52,59]	16 [15,18]	91 [89,93]	26 [24,28]	53 [51,55]
FLARE23 Top 1	18 [15,22]	100 [100,100]	31 [26,36]	59 [57,61]	8 [7,10]	100 [99,100]	15 [13,16]	54 [53,55]
FLARE23 Top 2	36 [32,40]	97 [94,100]	53 [48,57]	67 [64,69]	18 [17,20]	99 [98,99]	28 [26,29]	58 [57,59]
FLARE24 Top 1	22 [18,26]	97 [94,100]	36 [31,41]	60 [57,62]	13 [12,14]	99 [99,100]	22 [20,23]	56 [55,57]
FLARE25 Baseline	20 [16,24]	99 [97,100]	33 [28,38]	60 [57,62]	18 [17,20]	97 [95,98]	29 [27,31]	58 [57,59]
segmentation from reports (ours)								
R-Super	76 [72,80]	92 [87,97]	86 [83,88]	88 [85,91]	78 [77,80]	84 [81,87]	85 [84,86]	86 [84,88]

Table D17: **Tumor detection, external validation (Merlin, Stanford, USA): trained with no mask, R-Super surpasses state-of-the-art public AI models in the detection of 7 malignant tumor types.** This table evaluates tumor detection (malignant + benign vs no-tumor). R-Super was trained with the UCSF dataset (no tumor mask), and it was evaluated on the Merlin test set (Stanford hospital). No data from the Stanford hospital was used in training, making this an external (out-of-distribution) test set. See dataset details in Tab. 1, A1. We compare R-Super to prominent public AI models: MedGemma, Merlin, MedImageInsight, Lingshu, ULS, and winners of the FLARE cancer segmentation challenge. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best, using DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples.

AI	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	69 [63,75]	51 [41,60]	70 [65,75]	60 [53,66]	10 [2,19]	99 [97,100]	17 [4,32]	50 [40,61]
Merlin	59 [53,66]	84 [77,90]	71 [65,76]	76 [71,81]	63 [49,79]	79 [71,86]	57 [44,68]	68 [56,81]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,51]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	39 [33,46]	74 [66,82]	51 [44,57]	51 [45,57]	7 [0,16]	99 [97,100]	13 [0,27]	50 [44,56]
FLARE23 Top 1	3 [1,6]	100 [100,100]	6 [2,12]	52 [51,53]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	12 [3,23]	100 [100,100]	22 [6,37]	56 [52,61]
FLARE24 Top 1	8 [4,12]	99 [97,100]	14 [8,21]	53 [51,55]	2 [0,9]	100 [100,100]	5 [0,16]	51 [50,54]
FLARE25 Baseline	14 [10,19]	100 [100,100]	25 [18,32]	57 [55,60]	2 [0,8]	100 [100,100]	5 [0,15]	51 [50,54]
segmentation from reports (ours)								
R-Super	75 [69,81]	82 [73,88]	81 [76,85]	82 [77,87]	76 [62,88]	69 [61,78]	58 [46,69]	79 [70,87]

AI	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	30 [24,37]	89 [83,94]	44 [37,52]	62 [56,68]	78 [73,82]	29 [21,37]	77 [74,81]	50 [44,57]
Merlin	52 [46,60]	69 [61,78]	62 [55,68]	63 [57,69]	62 [57,67]	81 [74,88]	74 [70,78]	75 [70,80]
VLM report generation								
Merlin	1 [0,3]	100 [100,100]	2 [0,5]	50 [50,51]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]
Lingshu	4 [2,7]	98 [96,100]	8 [3,13]	51 [49,53]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	24 [18,30]	92 [87,97]	37 [29,44]	58 [54,61]	55 [50,61]	68 [60,77]	67 [62,71]	63 [58,68]
FLARE23 Top 1	18 [13,23]	100 [100,100]	30 [23,38]	59 [56,62]	6 [3,8]	100 [100,100]	11 [7,15]	53 [52,54]
FLARE23 Top 2	32 [25,38]	98 [96,100]	48 [40,55]	65 [61,68]	6 [4,9]	99 [97,100]	11 [7,16]	52 [51,54]
FLARE24 Top 1	19 [14,25]	100 [100,100]	32 [24,40]	60 [57,62]	22 [17,26]	99 [97,100]	36 [29,41]	60 [58,63]
FLARE25 Baseline	16 [11,21]	100 [100,100]	28 [20,35]	58 [56,61]	43 [37,48]	87 [81,93]	58 [53,63]	66 [62,70]
segmentation from reports (ours)								
R-Super	74 [69,80]	75 [67,83]	79 [75,84]	80 [75,85]	76 [71,80]	85 [78,91]	84 [81,87]	90 [87,93]

AI	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	34 [24,43]	74 [66,82]	42 [32,51]	52 [44,59]	16 [13,18]	88 [82,93]	27 [23,30]	45 [40,50]
Merlin	62 [52,71]	76 [68,84]	66 [58,74]	75 [68,82]	59 [55,62]	85 [78,91]	73 [70,76]	75 [71,79]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	1 [0,1]	100 [100,100]	2 [0,3]	50 [50,51]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	49 [46,51]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	60 [50,69]	74 [66,81]	64 [56,71]	69 [62,75]	2 [2,4]	100 [100,100]	5 [3,7]	51 [51,52]
FLARE23 Top 1	1 [0,3]	100 [100,100]	2 [0,6]	50 [50,52]	12 [10,15]	100 [100,100]	22 [19,26]	56 [55,57]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]	22 [19,25]	100 [100,100]	36 [32,40]	61 [60,63]
FLARE24 Top 1	5 [1,9]	99 [97,100]	9 [2,17]	52 [50,54]	14 [12,17]	100 [100,100]	25 [21,28]	57 [56,58]
FLARE25 Baseline	30 [22,39]	87 [81,93]	42 [32,51]	59 [54,64]	11 [9,13]	100 [100,100]	20 [17,24]	56 [55,57]
segmentation from reports (ours)								
R-Super	76 [68,84]	90 [84,95]	82 [75,87]	89 [84,92]	84 [81,87]	91 [86,96]	91 [89,92]	93 [90,95]

AI	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	92 [90,95]	18 [11,26]	87 [85,90]	56 [50,62]	47 [44,49]	64 [61,67]	52 [49,55]	54 [50,57]
Merlin	54 [50,59]	88 [82,93]	69 [65,73]	71 [66,76]	59 [56,62]	80 [75,86]	67 [64,71]	72 [68,76]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]	0 [0,0]	100 [100,100]	0 [0,1]	50 [50,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	49 [46,53]	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]
Lingshu	27 [23,31]	74 [66,82]	40 [35,45]	50 [46,55]	4 [4,5]	96 [95,97]	7 [6,8]	50 [49,51]
segmentation from masks								
ULS	38 [34,42]	78 [71,85]	53 [49,58]	59 [55,63]	32 [30,35]	84 [81,86]	41 [38,44]	57 [55,60]
FLARE23 Top 1	15 [12,18]	99 [97,100]	25 [20,30]	57 [55,59]	8 [7,9]	100 [100,100]	14 [12,16]	54 [53,54]
FLARE23 Top 2	29 [25,33]	96 [92,99]	44 [40,49]	62 [60,65]	14 [13,16]	99 [98,100]	23 [20,26]	57 [56,58]
FLARE24 Top 1	19 [15,22]	100 [100,100]	32 [27,36]	59 [58,61]	13 [11,14]	100 [99,100]	22 [19,24]	56 [55,57]
FLARE25 Baseline	14 [11,17]	100 [100,100]	24 [20,29]	57 [55,58]	19 [17,21]	96 [95,98]	29 [26,32]	58 [56,59]
segmentation from reports (ours)								
R-Super	68 [64,72]	90 [85,95]	80 [77,83]	81 [77,84]	76 [73,78]	83 [81,86]	79 [77,81]	85 [83,86]

Table D18: **Tumor detection, external validation: R-Super generalized well to European images (Medipol, Istanbul, Turkey) despite being trained only on American images (UCSF and Stanford, USA).** This table evaluates tumor detection (malignant + benign vs no-tumor). R-Super was trained with the R-Super train dataset (USA), and it was evaluated on the Medipol test set (external validation, Turkey). See dataset details in Tab. 1, A1. We compare R-Super to prominent public AI models: MedGemma, Merlin, MedImageInsight, Lingshu, ULS, and winners of the FLARE cancer segmentation challenge. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best, using DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples.

AI	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	51 [43,59]	79 [72,84]	57 [50,64]	71 [66,76]	8 [4,12]	94 [90,97]	13 [7,20]	46 [40,52]
Merlin	54 [46,62]	74 [68,80]	58 [51,65]	67 [60,72]	76 [70,83]	48 [41,56]	68 [62,73]	67 [62,72]
VLM report generation								
Merlin	6 [2,9]	99 [98,100]	10 [4,17]	52 [51,55]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	1 [0,4]	99 [98,100]	3 [0,7]	50 [49,52]	2 [0,4]	99 [98,100]	3 [0,7]	50 [49,52]
Lingshu	3 [1,7]	98 [96,100]	6 [1,13]	51 [49,53]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	62 [54,69]	72 [66,80]	63 [56,70]	70 [63,75]	37 [30,45]	86 [82,91]	50 [42,57]	62 [57,67]
FLARE23 Top 1	1 [0,2]	99 [98,100]	1 [0,4]	50 [48,51]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	17 [12,22]	100 [100,100]	29 [21,36]	59 [56,61]
FLARE24 Top 1	1 [0,2]	100 [100,100]	1 [0,4]	50 [50,51]	8 [4,11]	100 [100,100]	14 [8,20]	54 [52,56]
FLARE25 Baseline	7 [3,11]	99 [98,100]	13 [6,20]	53 [51,55]	6 [3,10]	98 [96,100]	12 [6,18]	52 [50,54]
segmentation from reports (ours)								
R-Super	90 [84,94]	65 [58,71]	77 [72,82]	84 [80,88]	77 [71,83]	89 [84,93]	82 [77,86]	89 [86,92]

AI	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	48 [40,56]	82 [76,87]	57 [50,64]	67 [61,73]	92 [86,96]	24 [18,30]	62 [56,67]	58 [52,64]
Merlin	54 [46,62]	69 [63,76]	58 [52,65]	64 [58,70]	56 [48,64]	82 [76,87]	62 [56,69]	71 [65,76]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	49 [48,50]	30 [23,38]	87 [82,92]	41 [32,49]	59 [54,64]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,51]	2 [0,4]	100 [100,100]	3 [0,8]	51 [50,52]
Lingshu	2 [1,5]	100 [100,100]	4 [1,10]	51 [50,52]	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]
segmentation from masks								
ULS	47 [39,54]	82 [76,87]	56 [49,63]	65 [60,70]	77 [70,85]	74 [68,80]	73 [66,78]	79 [74,84]
FLARE23 Top 1	3 [1,6]	98 [96,100]	6 [1,10]	51 [49,52]	9 [4,15]	99 [98,100]	17 [8,26]	54 [51,57]
FLARE23 Top 2	10 [6,15]	97 [94,99]	18 [11,25]	54 [51,56]	11 [6,16]	100 [100,100]	19 [11,27]	55 [52,58]
FLARE24 Top 1	3 [1,6]	99 [98,100]	7 [2,12]	51 [50,53]	35 [26,43]	97 [94,99]	50 [40,58]	66 [62,70]
FLARE25 Baseline	3 [1,6]	98 [96,100]	7 [2,12]	50 [48,52]	48 [39,57]	94 [90,97]	61 [53,69]	72 [67,76]
segmentation from reports (ours)								
R-Super	60 [52,67]	68 [62,75]	62 [56,68]	68 [63,74]	88 [82,93]	76 [70,82]	80 [74,84]	88 [84,92]

AI	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	66 [50,81]	62 [55,70]	37 [26,47]	63 [54,72]	68 [63,72]	49 [42,56]	70 [66,74]	58 [53,63]
Merlin	77 [62,90]	73 [66,79]	49 [37,60]	76 [66,84]	61 [57,66]	80 [74,85]	72 [67,75]	75 [71,80]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	3 [2,5]	99 [97,100]	6 [3,10]	51 [50,52]
MedGemma	3 [0,10]	99 [97,100]	5 [0,17]	51 [49,54]	3 [1,5]	98 [95,100]	6 [2,9]	50 [49,52]
Lingshu	17 [5,31]	96 [93,99]	25 [8,41]	57 [51,64]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	83 [70,94]	75 [69,81]	53 [42,64]	82 [74,88]	15 [11,18]	97 [94,99]	25 [19,30]	56 [53,58]
FLARE23 Top 1	3 [0,9]	99 [97,100]	5 [0,16]	51 [48,54]	6 [4,8]	100 [100,100]	11 [7,15]	53 [52,54]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]	13 [10,17]	100 [100,100]	23 [18,29]	56 [55,58]
FLARE24 Top 1	0 [0,0]	100 [100,100]	0 [0,0]	49 [47,50]	1 [0,3]	100 [100,100]	3 [1,5]	51 [50,51]
FLARE25 Baseline	17 [5,31]	96 [92,98]	24 [8,41]	56 [50,63]	1 [0,2]	100 [100,100]	2 [0,4]	51 [50,51]
segmentation from reports (ours)								
R-Super	80 [67,93]	97 [94,99]	81 [70,90]	93 [86,98]	88 [84,91]	91 [87,95]	92 [89,94]	94 [91,96]

AI	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	77 [72,81]	37 [30,45]	74 [70,77]	55 [49,60]	58 [55,61]	61 [58,64]	53 [50,56]	60 [56,63]
Merlin	68 [64,73]	65 [58,72]	74 [70,77]	68 [62,72]	64 [61,67]	70 [65,76]	63 [60,66]	70 [66,74]
VLM report generation								
Merlin	3 [1,5]	99 [97,100]	5 [2,9]	51 [50,52]	6 [5,7]	98 [97,99]	9 [7,11]	52 [51,53]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	49 [47,51]	2 [1,3]	99 [99,100]	3 [1,5]	50 [50,51]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	50 [46,54]	3 [2,5]	99 [99,100]	5 [2,8]	51 [50,52]
segmentation from masks								
ULS	66 [61,71]	51 [43,58]	69 [65,73]	58 [54,63]	55 [52,58]	77 [74,80]	56 [52,58]	67 [65,69]
FLARE23 Top 1	28 [23,33]	97 [94,99]	43 [37,49]	62 [60,65]	7 [6,9]	99 [98,100]	12 [10,14]	53 [52,54]
FLARE23 Top 2	38 [33,43]	90 [86,94]	53 [48,58]	64 [60,67]	13 [11,14]	98 [98,99]	20 [18,22]	55 [54,56]
FLARE24 Top 1	22 [18,26]	95 [92,98]	36 [30,41]	59 [56,61]	10 [8,11]	99 [98,99]	16 [14,18]	54 [53,55]
FLARE25 Baseline	20 [16,24]	96 [92,98]	32 [27,38]	58 [55,60]	15 [12,17]	97 [96,98]	22 [18,25]	56 [55,57]
segmentation from reports (ours)								
R-Super	62 [56,67]	89 [84,94]	74 [70,78]	81 [77,84]	78 [75,80]	82 [80,84]	78 [76,81]	85 [84,87]

Table D19: **Tumor detection, external validation: R-Super generalized well to European images (Basel, Switzerland) despite being trained only on American images (UCSF and Stanford, USA).** This table evaluates tumor detection (malignant + benign vs no-tumor). R-Super was trained with the R-Super train dataset (USA), and it was evaluated on the Basel test set (external validation, Switzerland). See dataset details in Tab. 1, A1. We compare R-Super to prominent public AI models: MedGemma, Merlin, MedImageInsight, Lingshu, ULS, and winners of the FLARE cancer segmentation challenge. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best, using DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples.

AI	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	75 [70,80]	78 [66,89]	84 [80,87]	82 [76,87]	99 [96,100]	17 [6,27]	79 [72,84]	46 [36,57]
Merlin	22 [18,27]	93 [85,98]	36 [30,42]	52 [44,59]	31 [22,42]	76 [64,86]	42 [31,53]	45 [36,54]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	49 [47,51]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	2 [1,4]	100 [100,100]	4 [1,7]	51 [50,52]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Lingshu	9 [6,12]	93 [85,98]	16 [10,22]	51 [46,54]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	83 [79,88]	65 [52,78]	88 [85,91]	73 [66,81]	66 [56,76]	63 [50,77]	70 [61,77]	66 [57,75]
FLARE23 Top 1	2 [1,4]	100 [100,100]	4 [1,7]	51 [50,52]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	12 [5,18]	100 [100,100]	21 [10,31]	56 [53,59]
FLARE24 Top 1	5 [2,7]	100 [100,100]	9 [5,14]	52 [50,54]	7 [2,13]	100 [100,100]	13 [4,22]	53 [51,56]
FLARE25 Baseline	15 [11,20]	98 [94,100]	26 [20,32]	57 [54,59]	17 [10,25]	100 [100,100]	29 [18,40]	58 [54,62]
segmentation from reports (ours)								
R-Super	71 [65,76]	92 [84,98]	82 [78,86]	89 [84,92]	85 [77,92]	89 [80,96]	89 [83,93]	90 [84,94]

AI	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	72 [62,82]	65 [52,77]	73 [65,80]	69 [59,78]	80 [75,85]	32 [20,45]	82 [79,86]	48 [39,57]
Merlin	59 [47,70]	70 [58,83]	66 [55,74]	64 [54,74]	36 [30,42]	87 [78,95]	52 [46,58]	59 [52,66]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	49 [46,52]	24 [19,29]	93 [85,98]	38 [31,45]	58 [54,62]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Lingshu	7 [2,13]	96 [90,100]	12 [3,22]	51 [48,56]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	50 [39,63]	76 [64,87]	60 [49,70]	64 [54,73]	67 [61,72]	94 [88,100]	80 [75,84]	84 [80,88]
FLARE23 Top 1	14 [7,23]	98 [94,100]	25 [12,37]	55 [50,60]	5 [2,7]	100 [100,100]	9 [4,14]	52 [51,54]
FLARE23 Top 2	17 [9,26]	96 [91,100]	29 [16,40]	57 [51,63]	4 [2,6]	100 [100,100]	7 [3,12]	52 [51,53]
FLARE24 Top 1	16 [8,25]	100 [100,100]	27 [15,40]	58 [54,62]	19 [14,24]	100 [100,100]	32 [25,38]	60 [57,62]
FLARE25 Baseline	8 [2,15]	100 [100,100]	15 [5,26]	53 [49,57]	31 [26,37]	100 [100,100]	48 [41,54]	66 [63,68]
segmentation from reports (ours)								
R-Super	70 [59,80]	53 [39,68]	69 [60,77]	64 [55,75]	64 [58,70]	100 [100,100]	78 [73,82]	83 [78,87]

AI	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	51 [43,58]	70 [57,82]	63 [57,69]	60 [51,69]	23 [21,25]	87 [76,95]	37 [34,40]	46 [39,53]
Merlin	8 [4,13]	100 [100,100]	15 [9,22]	46 [38,54]	21 [19,23]	89 [80,96]	34 [32,37]	45 [39,53]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	48 [45,50]	0 [0,0]	100 [100,100]	0 [0,0]	49 [46,51]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,1]	100 [100,100]	1 [0,2]	50 [50,50]
Lingshu	15 [10,20]	96 [90,100]	26 [18,33]	56 [52,59]	0 [0,0]	100 [100,100]	0 [0,1]	50 [50,50]
segmentation from masks								
ULS	64 [56,71]	87 [77,96]	76 [70,81]	78 [72,83]	28 [26,30]	96 [91,100]	44 [41,46]	60 [55,65]
FLARE23 Top 1	2 [0,4]	100 [100,100]	4 [0,7]	51 [50,52]	11 [10,13]	98 [94,100]	20 [17,23]	55 [52,56]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	20 [18,22]	100 [100,100]	33 [31,36]	60 [59,61]
FLARE24 Top 1	2 [0,4]	100 [100,100]	4 [0,8]	51 [50,52]	10 [9,12]	100 [100,100]	18 [16,21]	55 [54,56]
FLARE25 Baseline	14 [9,20]	100 [100,100]	25 [17,33]	57 [55,60]	8 [7,10]	100 [100,100]	15 [13,18]	54 [53,55]
segmentation from reports (ours)								
R-Super	82 [76,88]	89 [79,96]	89 [84,92]	88 [83,93]	72 [69,74]	85 [74,94]	83 [82,85]	83 [78,88]

AI	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	91 [88,93]	17 [7,26]	92 [90,93]	44 [36,52]	70 [68,72]	52 [47,57]	73 [70,75]	57 [52,61]
Merlin	18 [15,21]	91 [82,98]	30 [26,34]	47 [40,53]	28 [25,31]	86 [80,92]	39 [36,42]	51 [45,57]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [47,52]	3 [3,4]	99 [98,100]	5 [4,6]	50 [48,52]
MedGemma	0 [0,1]	100 [100,100]	1 [0,2]	50 [50,50]	0 [0,1]	100 [100,100]	1 [0,1]	50 [50,50]
Lingshu	23 [19,26]	80 [68,90]	36 [32,40]	51 [45,57]	8 [6,9]	95 [93,97]	13 [11,15]	51 [50,53]
segmentation from masks								
ULS	81 [78,84]	17 [7,27]	86 [84,88]	34 [25,44]	63 [60,65]	71 [67,75]	72 [70,74]	66 [63,68]
FLARE23 Top 1	30 [27,34]	94 [88,100]	47 [43,51]	63 [59,66]	9 [8,11]	99 [97,100]	16 [13,17]	54 [53,55]
FLARE23 Top 2	49 [45,52]	85 [76,94]	65 [62,68]	68 [63,73]	14 [13,16]	97 [96,99]	22 [20,25]	56 [55,57]
FLARE24 Top 1	28 [24,31]	82 [71,92]	43 [39,46]	53 [47,59]	12 [11,14]	97 [96,99]	21 [18,23]	55 [53,56]
FLARE25 Baseline	20 [17,23]	85 [75,94]	34 [29,38]	51 [44,57]	16 [14,18]	98 [96,99]	27 [24,30]	56 [55,58]
segmentation from reports (ours)								
R-Super	67 [64,70]	85 [75,94]	80 [77,82]	82 [77,86]	73 [70,75]	85 [81,88]	81 [79,83]	83 [80,85]

D.6.2 Malignant Tumor Detection

Table D20: **Malignant tumors, internal validation (UCSF, USA): trained with no mask, R-Super surpasses state-of-the-art public AI models in the detection of 7 malignant tumor types.** This table evaluates tumor detection stratified for malignant tumors (malignant vs no-tumor). R-Super was trained with the R-Super train set (USA, no tumor mask), and it was evaluated on the UCSF test set (internal validation). See dataset details in Tab. 1, A1. We compare R-Super to prominent public AI models: MedGemma, Merlin, MedImageInsight, Lingshu, ULS, and winners of the FLARE cancer segmentation challenge. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model for each metric. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples.

AI	bladder malignant tumor				esophagus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	31 [25,38]	78 [70,86]	44 [37,51]	53 [46,59]	44 [38,50]	80 [73,88]	57 [50,63]	62 [56,68]
Merlin	42 [35,49]	81 [74,88]	55 [48,61]	65 [59,71]	60 [54,66]	69 [60,77]	69 [64,74]	70 [64,75]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,51]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	49 [48,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	26 [20,33]	79 [71,86]	37 [30,45]	51 [45,57]	5 [3,8]	98 [95,100]	10 [5,15]	51 [47,54]
FLARE23 Top 1	7 [4,11]	99 [97,100]	13 [7,19]	53 [50,55]	1 [0,2]	100 [100,100]	2 [0,4]	50 [50,51]
FLARE23 Top 2	3 [1,6]	99 [97,100]	6 [2,11]	51 [50,53]	12 [8,16]	99 [97,100]	20 [14,27]	55 [53,57]
FLARE24 Top 1	8 [4,12]	99 [97,100]	14 [8,21]	53 [51,55]	4 [2,6]	100 [100,100]	7 [3,12]	52 [51,53]
FLARE25 Baseline	11 [7,16]	97 [94,100]	20 [12,27]	54 [52,57]	7 [4,11]	100 [100,100]	14 [8,19]	54 [52,55]
segmentation from reports (ours)								
R-Super	77 [71,82]	78 [70,85]	81 [76,85]	82 [77,86]	88 [83,92]	87 [80,93]	90 [87,93]	92 [89,95]

AI	gallbladder malignant tumor				uterus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	83 [75,91]	42 [34,52]	62 [55,70]	65 [57,72]	57 [46,68]	57 [48,65]	53 [44,61]	55 [47,63]
Merlin	31 [21,41]	80 [73,88]	39 [27,48]	52 [44,61]	71 [61,81]	78 [70,85]	71 [62,79]	76 [68,84]
VLM report generation								
Merlin	1 [0,4]	100 [100,100]	2 [0,8]	51 [50,52]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	4 [0,9]	100 [100,100]	8 [0,17]	52 [50,55]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	48 [46,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	13 [6,21]	96 [92,99]	22 [11,33]	54 [49,58]	32 [23,42]	85 [78,91]	42 [31,53]	58 [52,65]
FLARE23 Top 1	35 [24,46]	98 [95,100]	50 [37,62]	66 [61,72]	11 [5,17]	100 [100,100]	19 [9,30]	55 [52,59]
FLARE23 Top 2	83 [74,91]	94 [89,98]	87 [80,92]	89 [84,94]	17 [10,25]	99 [97,100]	28 [17,40]	58 [54,62]
FLARE24 Top 1	55 [44,66]	98 [95,100]	70 [60,78]	77 [71,82]	52 [41,63]	98 [96,100]	68 [57,77]	76 [70,81]
FLARE25 Baseline	50 [39,61]	98 [95,100]	66 [55,75]	74 [68,80]	61 [50,70]	92 [87,97]	71 [62,79]	77 [71,83]
segmentation from reports (ours)								
R-Super	90 [82,96]	93 [88,97]	90 [85,94]	95 [91,99]	80 [71,88]	96 [92,99]	86 [80,91]	91 [86,95]

AI	prostate malignant tumor				adrenal malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	58 [49,69]	58 [50,67]	55 [47,64]	56 [49,64]	32 [23,39]	77 [70,84]	42 [32,50]	52 [45,59]
Merlin	64 [54,74]	81 [74,88]	68 [60,76]	75 [68,82]	77 [70,84]	46 [37,56]	69 [62,75]	65 [58,72]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	4 [1,8]	98 [96,100]	9 [3,15]	51 [49,54]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	33 [24,42]	85 [78,91]	43 [32,53]	58 [52,65]	1 [0,3]	100 [100,100]	2 [0,5]	50 [50,51]
FLARE23 Top 1	1 [0,4]	100 [100,100]	2 [0,7]	51 [50,52]	30 [23,38]	100 [100,100]	46 [37,55]	65 [61,69]
FLARE23 Top 2	1 [0,4]	100 [100,100]	2 [0,7]	50 [49,52]	35 [27,44]	100 [100,100]	52 [43,61]	68 [64,72]
FLARE24 Top 1	2 [0,6]	100 [100,100]	4 [0,10]	50 [48,52]	19 [12,25]	100 [100,100]	32 [22,40]	59 [56,63]
FLARE25 Baseline	28 [20,38]	93 [88,97]	41 [30,51]	60 [55,65]	20 [13,26]	100 [100,100]	33 [23,42]	60 [56,63]
segmentation from reports (ours)								
R-Super	91 [85,96]	72 [63,80]	80 [73,86]	88 [82,93]	92 [87,96]	93 [88,97]	93 [90,96]	96 [94,98]

AI	spleen malignant tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	85 [72,97]	58 [50,67]	53 [42,64]	71 [62,79]	56 [52,59]	64 [60,69]	52 [48,56]	59 [56,63]
Merlin	59 [42,74]	74 [66,81]	48 [34,60]	71 [61,81]	58 [54,62]	73 [66,79]	60 [56,64]	68 [63,72]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]	0 [0,1]	100 [100,100]	0 [0,1]	50 [50,50]
MedGemma	6 [0,16]	95 [90,98]	10 [0,23]	50 [46,55]	2 [1,4]	99 [98,100]	4 [2,6]	50 [50,51]
Lingshu	24 [10,39]	77 [69,84]	24 [11,37]	50 [42,59]	3 [2,6]	97 [96,98]	3 [2,5]	50 [48,51]
segmentation from masks								
ULS	28 [13,44]	90 [84,96]	35 [17,50]	58 [50,67]	20 [16,23]	90 [88,93]	27 [23,31]	54 [52,57]
FLARE23 Top 1	15 [3,29]	100 [100,100]	26 [6,44]	57 [51,64]	14 [11,17]	100 [99,100]	23 [18,26]	57 [55,58]
FLARE23 Top 2	44 [27,62]	98 [95,100]	59 [40,75]	71 [62,80]	28 [25,31]	98 [98,99]	36 [33,40]	63 [61,65]
FLARE24 Top 1	18 [6,31]	98 [95,100]	29 [11,46]	57 [52,64]	23 [19,25]	99 [98,100]	32 [28,35]	61 [59,62]
FLARE25 Baseline	24 [11,39]	99 [97,100]	37 [19,55]	61 [55,69]	29 [25,32]	97 [95,98]	40 [36,44]	63 [61,65]
segmentation from reports (ours)								
R-Super	91 [81,100]	92 [87,97]	84 [74,92]	95 [90,98]	87 [84,90]	87 [84,90]	86 [84,88]	91 [89,93]

Table D21: **Malignant tumors, external validation (Merlin, Stanford, USA): trained with no mask, R-Super surpasses state-of-the-art public AI models in the detection of 7 malignant tumor types.** This table evaluates tumor detection stratified for malignant tumors (malignant vs no-tumor). R-Super was trained with the UCSF dataset (no tumor mask), and it was evaluated on the Merlin test set (Stanford hospital). No data from the Stanford hospital was used in training, making this an external (out-of-distribution) test set. See dataset details in Tab. 1, A1. We compare R-Super to prominent public AI models: MedGemma, Merlin, MedImageInsight, Lingshu, ULS, and winners of the FLARE cancer segmentation challenge. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best, using DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples.

AI	bladder malignant tumor				esophagus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	64 [50,76]	54 [43,63]	49 [38,59]	55 [46,65]	13 [0,33]	95 [91,98]	17 [0,38]	41 [25,58]
Merlin	71 [59,82]	91 [86,96]	75 [65,83]	86 [80,91]	60 [31,85]	90 [85,96]	51 [27,71]	62 [37,84]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	49 [48,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	44 [31,56]	75 [66,83]	44 [32,56]	54 [44,63]	13 [0,35]	99 [97,100]	22 [0,48]	51 [43,63]
FLARE23 Top 1	4 [0,9]	100 [100,100]	7 [0,17]	52 [50,55]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	13 [0,33]	100 [100,100]	24 [0,50]	57 [50,67]
FLARE24 Top 1	11 [4,19]	99 [97,100]	19 [7,32]	55 [51,59]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE25 Baseline	26 [14,37]	100 [100,100]	41 [24,54]	63 [57,69]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from reports (ours)								
R-Super	82 [71,91]	83 [76,89]	76 [66,84]	84 [76,91]	73 [50,93]	89 [82,94]	56 [36,74]	83 [68,94]

AI	gallbladder malignant tumor				uterus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	98 [92,100]	23 [16,31]	47 [38,56]	62 [53,71]	93 [86,98]	12 [7,19]	50 [42,59]	46 [36,55]
Merlin	93 [83,100]	29 [21,38]	48 [38,56]	63 [52,72]	81 [70,90]	87 [80,93]	78 [69,86]	85 [78,92]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]
Lingshu	5 [0,13]	98 [96,100]	9 [0,22]	52 [49,56]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	34 [20,48]	92 [87,97]	44 [29,58]	63 [56,71]	72 [60,82]	69 [61,78]	62 [51,71]	72 [64,79]
FLARE23 Top 1	58 [44,75]	99 [97,100]	73 [60,85]	79 [72,87]	9 [2,18]	100 [100,100]	16 [4,30]	54 [51,59]
FLARE23 Top 2	71 [56,84]	98 [96,100]	81 [70,89]	85 [77,91]	4 [0,9]	99 [97,100]	7 [0,16]	51 [49,54]
FLARE24 Top 1	58 [42,74]	100 [100,100]	74 [60,85]	79 [71,87]	46 [34,60]	99 [97,100]	62 [50,74]	73 [66,80]
FLARE25 Baseline	51 [35,66]	100 [100,100]	68 [52,80]	76 [68,83]	63 [50,75]	87 [81,93]	67 [56,76]	77 [70,84]
segmentation from reports (ours)								
R-Super	76 [62,89]	96 [92,99]	80 [70,90]	92 [86,97]	81 [70,91]	93 [88,97]	83 [74,90]	95 [92,98]

AI	prostate malignant tumor				adrenal malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	14 [0,31]	92 [87,97]	18 [0,36]	39 [26,54]	5 [1,11]	96 [93,99]	10 [2,18]	43 [34,51]
Merlin	71 [50,91]	92 [87,96]	67 [48,80]	88 [79,95]	79 [70,86]	92 [87,97]	84 [77,88]	87 [81,92]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	2 [0,5]	100 [100,100]	4 [0,10]	51 [50,53]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	48 [46,50]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	67 [45,86]	84 [77,91]	53 [35,67]	79 [67,89]	5 [1,10]	100 [100,100]	10 [2,18]	53 [51,55]
FLARE23 Top 1	5 [0,15]	100 [100,100]	9 [0,27]	52 [50,58]	27 [18,36]	100 [100,100]	42 [30,53]	63 [59,68]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]	48 [38,58]	100 [100,100]	65 [55,74]	74 [69,79]
FLARE24 Top 1	19 [4,36]	99 [97,100]	31 [8,50]	59 [51,68]	28 [20,37]	100 [100,100]	43 [33,54]	64 [60,68]
FLARE25 Baseline	57 [36,78]	87 [81,93]	50 [33,65]	71 [60,81]	27 [19,36]	100 [100,100]	42 [32,52]	63 [59,68]
segmentation from reports (ours)								
R-Super	90 [76,100]	90 [84,95]	74 [60,86]	95 [89,99]	95 [90,99]	96 [92,99]	95 [91,98]	98 [95,100]

AI	spleen malignant tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	96 [90,100]	25 [18,34]	60 [52,68]	60 [52,69]	55 [50,59]	57 [54,60]	36 [30,42]	50 [45,54]
Merlin	72 [62,83]	88 [82,93]	75 [66,83]	79 [71,87]	75 [69,81]	81 [77,85]	68 [62,74]	78 [73,83]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]	0 [0,1]	100 [100,100]	1 [0,2]	50 [50,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	49 [44,54]	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]
Lingshu	33 [22,46]	74 [66,82]	38 [25,49]	54 [46,61]	6 [4,8]	96 [95,97]	7 [4,9]	51 [49,52]
segmentation from masks								
ULS	48 [36,60]	96 [92,99]	62 [49,72]	74 [67,81]	40 [35,46]	88 [86,90]	42 [37,48]	64 [60,67]
FLARE23 Top 1	35 [23,46]	99 [97,100]	51 [37,62]	67 [61,73]	20 [16,23]	100 [99,100]	28 [24,33]	60 [58,62]
FLARE23 Top 2	59 [49,70]	96 [92,99]	71 [62,80]	78 [72,84]	28 [24,32]	99 [98,100]	35 [31,40]	63 [61,66]
FLARE24 Top 1	30 [20,41]	100 [100,100]	47 [33,58]	65 [60,70]	28 [23,32]	100 [99,100]	39 [34,44]	64 [61,66]
FLARE25 Baseline	32 [22,42]	100 [100,100]	48 [36,60]	66 [61,71]	36 [31,42]	96 [95,98]	45 [39,50]	66 [64,69]
segmentation from reports (ours)								
R-Super	90 [83,96]	90 [84,95]	87 [80,92]	92 [88,96]	84 [79,88]	91 [89,93]	79 [74,83]	91 [89,94]

Table D22: **Malignant tumors, external validation: R-Super generalized well to European images (Medipol, Istanbul, Turkey) despite being trained only on American images (UCSF and Stanford, USA).** This table evaluates tumor detection stratified for malignant tumors (malignant vs no-tumor). R-Super was trained with the R-Super train dataset (USA), and it was evaluated on the Medipol test set (external validation, Turkey). See dataset details in Tab. 1, A1. We compare R-Super to prominent public AI models: MedGemma, Merlin, MedImageInsight, Lingshu, ULS, and winners of the FLARE cancer segmentation challenge. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best, using DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples. n/a: we skipped organs with less than 10 positive scans in the dataset.

AI	bladder malignant tumor				esophagus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	74 [60,87]	79 [72,84]	55 [43,65]	80 [73,88]	54 [41,67]	52 [45,60]	34 [26,43]	48 [39,57]
Merlin	49 [32,63]	79 [73,85]	40 [27,51]	64 [53,72]	37 [25,51]	86 [82,91]	41 [28,53]	66 [58,74]
VLM report generation								
Merlin	3 [0,8]	99 [98,100]	5 [0,15]	51 [49,54]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]	2 [0,6]	99 [98,100]	4 [0,11]	51 [49,53]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	49 [48,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	67 [52,82]	66 [58,73]	41 [30,52]	68 [57,78]	20 [10,32]	97 [94,99]	31 [17,45]	57 [50,64]
FLARE23 Top 1	3 [0,9]	99 [98,100]	5 [0,15]	50 [48,54]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	26 [15,38]	100 [100,100]	41 [26,55]	63 [57,69]
FLARE24 Top 1	3 [0,9]	100 [100,100]	5 [0,16]	51 [50,54]	15 [6,24]	100 [100,100]	26 [12,39]	57 [53,62]
FLARE25 Baseline	13 [3,24]	99 [98,100]	22 [5,38]	56 [51,62]	9 [2,18]	98 [96,100]	16 [4,29]	54 [50,58]
segmentation from reports (ours)								
R-Super	82 [70,94]	94 [90,97]	78 [67,87]	90 [83,96]	93 [85,98]	93 [88,96]	86 [78,92]	97 [93,99]

AI	gallbladder malignant tumor				uterus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Merlin	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
VLM report generation								
Merlin	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
MedGemma	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Lingshu	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
segmentation from masks								
ULS	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
FLARE23 Top 1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
FLARE23 Top 2	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
FLARE24 Top 1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
FLARE25 Baseline	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
segmentation from reports (ours)								
R-Super	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

AI	prostate malignant tumor				adrenal malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	n/a	n/a	n/a	n/a	75 [64,86]	49 [42,56]	45 [37,54]	60 [52,68]
Merlin	n/a	n/a	n/a	n/a	97 [92,100]	74 [67,80]	70 [62,78]	86 [81,90]
VLM report generation								
Merlin	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	49 [48,50]
MedGemma	n/a	n/a	n/a	n/a	5 [0,12]	98 [95,100]	9 [0,20]	51 [49,55]
Lingshu	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	n/a	n/a	n/a	n/a	12 [4,22]	100 [100,100]	21 [8,36]	54 [50,59]
FLARE23 Top 1	n/a	n/a	n/a	n/a	10 [3,19]	100 [100,100]	18 [6,32]	55 [52,60]
FLARE23 Top 2	n/a	n/a	n/a	n/a	36 [24,48]	100 [100,100]	52 [38,65]	68 [62,74]
FLARE24 Top 1	n/a	n/a	n/a	n/a	2 [0,6]	100 [100,100]	3 [0,11]	51 [50,53]
FLARE25 Baseline	n/a	n/a	n/a	n/a	2 [0,5]	100 [100,100]	3 [0,10]	51 [50,53]
segmentation from reports (ours)								
R-Super	n/a	n/a	n/a	n/a	95 [89,100]	91 [87,95]	86 [78,92]	97 [95,99]

AI	spleen malignant tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	68 [59,78]	48 [40,55]	53 [45,60]	57 [50,64]	68 [62,74]	57 [52,61]	47 [42,52]	61 [57,66]
Merlin	82 [74,89]	65 [58,72]	68 [61,74]	75 [69,81]	66 [60,72]	76 [71,81]	55 [48,61]	73 [68,77]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,51]	1 [0,2]	100 [100,100]	1 [0,4]	50 [49,51]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	49 [47,52]	2 [0,4]	99 [98,100]	3 [1,7]	50 [49,52]
Lingshu	24 [16,33]	78 [72,83]	30 [21,39]	51 [46,56]	6 [4,8]	94 [93,96]	7 [5,10]	50 [49,51]
segmentation from masks								
ULS	47 [38,57]	67 [61,74]	46 [38,55]	58 [51,65]	36 [31,42]	83 [80,85]	35 [29,41]	59 [55,63]
FLARE23 Top 1	46 [37,56]	98 [95,100]	62 [52,70]	72 [67,77]	15 [12,19]	99 [99,100]	21 [17,26]	57 [55,59]
FLARE23 Top 2	64 [55,72]	90 [86,94]	71 [63,78]	78 [73,83]	31 [26,37]	98 [96,99]	41 [35,47]	64 [62,67]
FLARE24 Top 1	54 [45,64]	95 [92,98]	66 [58,75]	75 [70,80]	18 [14,22]	99 [98,100]	25 [20,30]	59 [57,61]
FLARE25 Baseline	42 [33,52]	96 [92,98]	56 [46,65]	69 [64,74]	16 [13,21]	98 [97,99]	24 [19,30]	57 [56,60]
segmentation from reports (ours)								
R-Super	78 [69,85]	89 [84,94]	79 [73,85]	87 [82,92]	87 [82,91]	92 [89,94]	82 [77,86]	93 [90,95]

Table D23: **Malignant tumors, external validation: R-Super generalized well to European images (Basel, Switzerland) despite being trained only on American images (UCSF and Stanford, USA).** This table evaluates tumor detection stratified for malignant tumors (malignant vs no-tumor). R-Super was trained with the R-Super train dataset (USA), and it was evaluated on the Basel test set (external validation, Switzerland). See dataset details in Tab. 1, A1. We compare R-Super to prominent public AI models: MedGemma, Merlin, MedImageInsight, Lingshu, ULS, and winners of the FLARE cancer segmentation challenge. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best, using DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples.

AI	bladder malignant tumor				esophagus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	87 [80,94]	67 [53,79]	83 [76,89]	83 [76,90]	89 [76,100]	32 [20,44]	56 [42,67]	50 [38,64]
Merlin	39 [29,50]	78 [66,88]	51 [40,61]	53 [43,63]	43 [25,62]	76 [64,86]	45 [28,62]	54 [40,68]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	49 [47,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	2 [0,7]	100 [100,100]	5 [0,12]	51 [50,53]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Lingshu	8 [2,14]	93 [85,98]	14 [4,23]	50 [46,54]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	86 [77,93]	65 [52,78]	82 [75,88]	74 [65,83]	82 [67,95]	59 [46,72]	63 [48,75]	72 [61,83]
FLARE23 Top 1	1 [0,4]	100 [100,100]	2 [0,8]	51 [50,52]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	11 [0,24]	100 [100,100]	19 [0,39]	55 [50,62]
FLARE24 Top 1	10 [4,17]	98 [94,100]	18 [8,30]	54 [51,58]	4 [0,12]	100 [100,100]	7 [0,21]	52 [50,56]
FLARE25 Baseline	22 [12,31]	98 [94,100]	35 [22,47]	60 [55,65]	18 [4,32]	100 [100,100]	30 [7,49]	58 [51,66]
segmentation from reports (ours)								
R-Super	77 [68,86]	96 [90,100]	86 [80,92]	93 [88,97]	79 [61,92]	89 [80,96]	79 [64,88]	82 [69,92]

AI	gallbladder malignant tumor				uterus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	71 [48,91]	65 [52,77]	50 [32,66]	68 [52,81]	73 [54,89]	43 [28,56]	53 [38,65]	55 [42,68]
Merlin	65 [40,87]	70 [58,83]	50 [30,68]	64 [49,78]	87 [74,97]	68 [55,81]	71 [58,82]	81 [70,91]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	48 [45,50]	10 [0,23]	93 [85,98]	16 [0,33]	51 [46,58]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Lingshu	6 [0,22]	96 [90,100]	10 [0,32]	51 [46,59]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	88 [71,100]	63 [50,76]	58 [40,73]	75 [63,87]	100 [100,100]	96 [91,100]	97 [91,100]	100 [99,100]
FLARE23 Top 1	29 [8,53]	98 [94,100]	44 [12,67]	62 [51,75]	7 [0,17]	100 [100,100]	12 [0,29]	53 [50,58]
FLARE23 Top 2	65 [41,88]	91 [83,98]	67 [44,83]	78 [66,91]	7 [0,17]	100 [100,100]	12 [0,29]	53 [50,58]
FLARE24 Top 1	29 [8,53]	100 [100,100]	46 [15,69]	65 [54,76]	73 [56,89]	100 [100,100]	85 [72,94]	87 [78,94]
FLARE25 Baseline	18 [0,40]	96 [91,100]	27 [0,52]	57 [48,68]	73 [57,89]	100 [100,100]	85 [73,94]	87 [79,94]
segmentation from reports (ours)								
R-Super	65 [40,90]	76 [64,87]	54 [32,71]	69 [54,84]	93 [83,100]	98 [94,100]	95 [88,100]	95 [87,100]

AI	prostate malignant tumor				adrenal malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	79 [62,93]	52 [38,65]	58 [44,71]	64 [51,76]	26 [22,30]	87 [76,95]	40 [35,46]	50 [42,57]
Merlin	25 [10,42]	96 [90,100]	38 [17,56]	55 [41,69]	25 [21,30]	91 [83,98]	40 [35,45]	51 [44,58]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	48 [45,50]	0 [0,0]	100 [100,100]	0 [0,0]	49 [46,51]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	1 [0,2]	100 [100,100]	2 [0,4]	50 [50,51]
Lingshu	18 [4,32]	96 [90,100]	29 [7,46]	57 [50,64]	0 [0,1]	100 [100,100]	1 [0,2]	50 [50,51]
segmentation from masks								
ULS	68 [50,85]	83 [73,93]	68 [53,80]	79 [68,89]	30 [26,34]	96 [91,100]	46 [41,51]	60 [55,66]
FLARE23 Top 1	7 [0,18]	100 [100,100]	13 [0,30]	54 [50,59]	16 [13,20]	98 [94,100]	28 [23,33]	57 [55,60]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	31 [27,36]	100 [100,100]	48 [43,53]	66 [64,68]
FLARE24 Top 1	7 [0,18]	100 [100,100]	13 [0,30]	54 [50,59]	14 [11,18]	100 [100,100]	25 [20,30]	57 [56,59]
FLARE25 Baseline	14 [3,29]	100 [100,100]	25 [6,44]	57 [52,64]	11 [8,14]	100 [100,100]	20 [14,24]	56 [54,57]
segmentation from reports (ours)								
R-Super	82 [67,96]	89 [79,96]	81 [67,90]	87 [77,95]	72 [68,76]	85 [74,94]	83 [80,86]	83 [78,88]

AI	spleen malignant tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	81 [72,89]	33 [21,46]	74 [66,80]	53 [43,62]	72 [66,77]	54 [48,60]	59 [53,65]	60 [55,66]
Merlin	33 [24,43]	91 [82,98]	48 [36,58]	64 [54,72]	45 [40,51]	82 [74,88]	49 [43,55]	60 [54,67]
VLM report generation								
Merlin	6 [1,10]	96 [90,100]	10 [2,19]	51 [47,54]	2 [1,4]	98 [97,100]	4 [1,6]	50 [47,51]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,1]	100 [100,100]	1 [0,2]	50 [50,51]
Lingshu	22 [14,30]	80 [68,90]	33 [22,43]	51 [44,58]	8 [5,11]	95 [93,97]	12 [8,16]	51 [50,53]
segmentation from masks								
ULS	89 [82,95]	17 [7,27]	75 [68,81]	39 [28,50]	78 [73,82]	68 [65,72]	70 [65,74]	71 [68,75]
FLARE23 Top 1	47 [37,58]	94 [88,100]	63 [52,72]	72 [66,78]	15 [12,20]	99 [97,100]	23 [18,28]	57 [55,59]
FLARE23 Top 2	54 [43,64]	96 [90,100]	69 [59,77]	77 [69,83]	24 [20,28]	98 [96,100]	31 [26,35]	61 [59,64]
FLARE24 Top 1	40 [30,50]	85 [76,94]	53 [43,63]	61 [52,68]	25 [21,30]	98 [96,99]	35 [30,40]	61 [59,64]
FLARE25 Baseline	41 [29,51]	85 [75,94]	54 [42,64]	61 [53,69]	28 [23,33]	97 [96,98]	40 [33,45]	62 [60,65]
segmentation from reports (ours)								
R-Super	77 [67,86]	76 [63,86]	80 [73,86]	81 [74,88]	78 [73,83]	87 [83,90]	80 [75,83]	84 [81,87]

D.6.3 Small Tumor Detection

Table D24: **Small tumor detection, internal validation (UCSF, USA): trained with no mask, R-Super accurately detects small tumors (diameter ≤ 2 cm).** This table evaluates tumor detection (malignant + benign vs no-tumor), evaluating only on small tumors. R-Super was trained with the R-Super train set (USA, no tumor mask), and it was evaluated on the UCSF test set (internal validation). See dataset details in Tab. 1, A1. We compare R-Super to prominent public AI models: MedGemma, Merlin, MedImageInsight, Lingshu, ULS, and winners of the FLARE cancer segmentation challenge. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model for each metric. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples.

AI	bladder small tumor				esophagus small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	73 [65,81]	41 [32,50]	62 [55,69]	56 [49,64]	55 [42,67]	80 [73,88]	58 [47,68]	69 [61,77]
Merlin	36 [27,46]	81 [74,88]	47 [36,56]	59 [51,66]	45 [34,58]	69 [60,77]	45 [35,56]	60 [51,69]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,53]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,51]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	28 [20,37]	79 [71,86]	38 [28,47]	53 [46,60]	9 [3,17]	98 [95,100]	17 [5,29]	54 [48,60]
FLARE23 Top 1	4 [1,8]	99 [97,100]	7 [2,14]	51 [49,53]	2 [0,5]	100 [100,100]	3 [0,10]	51 [50,53]
FLARE23 Top 2	5 [1,9]	99 [97,100]	9 [2,17]	52 [50,54]	20 [11,30]	99 [97,100]	33 [20,46]	60 [55,65]
FLARE24 Top 1	3 [0,6]	99 [97,100]	5 [0,12]	50 [49,52]	5 [0,10]	100 [100,100]	9 [0,18]	52 [50,55]
FLARE25 Baseline	3 [0,6]	97 [94,100]	5 [0,12]	50 [48,52]	6 [1,13]	100 [100,100]	12 [3,23]	53 [51,56]
segmentation from reports (ours)								
R-Super	57 [48,66]	73 [64,80]	61 [54,69]	67 [60,74]	88 [79,95]	87 [80,93]	83 [75,89]	93 [89,96]

AI	gallbladder small tumor				uterus small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	62 [49,73]	44 [35,53]	48 [38,56]	52 [43,61]	50 [38,63]	57 [48,65]	43 [33,52]	53 [44,62]
Merlin	20 [11,31]	80 [73,88]	26 [15,38]	45 [36,54]	48 [36,61]	77 [69,84]	50 [39,62]	64 [54,73]
VLM report generation								
Merlin	2 [0,5]	100 [100,100]	3 [0,10]	51 [50,52]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	2 [0,5]	100 [100,100]	3 [0,10]	51 [50,53]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	51 [47,54]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	8 [2,14]	96 [92,99]	14 [3,24]	51 [47,55]	23 [12,34]	85 [78,91]	31 [17,43]	55 [47,62]
FLARE23 Top 1	11 [4,19]	98 [95,100]	19 [7,31]	55 [51,59]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE23 Top 2	25 [14,36]	95 [90,98]	37 [23,50]	59 [54,65]	2 [0,6]	100 [100,100]	3 [0,11]	50 [49,52]
FLARE24 Top 1	12 [5,22]	98 [95,100]	21 [9,35]	55 [51,60]	2 [0,6]	100 [100,100]	3 [0,10]	50 [48,52]
FLARE25 Baseline	12 [5,21]	98 [95,100]	21 [9,33]	55 [51,60]	20 [10,30]	92 [87,97]	30 [16,42]	55 [50,61]
segmentation from reports (ours)								
R-Super	55 [43,67]	72 [63,79]	54 [43,64]	70 [62,77]	75 [63,86]	87 [80,93]	75 [65,83]	86 [79,92]

AI	prostate small tumor				adrenal small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	68 [51,83]	54 [46,63]	40 [28,52]	61 [50,72]	67 [61,72]	45 [36,54]	71 [66,75]	55 [49,61]
Merlin	52 [32,69]	74 [65,81]	42 [26,55]	60 [47,72]	44 [39,50]	78 [70,85]	58 [52,63]	64 [58,70]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	4 [2,6]	98 [96,100]	8 [4,12]	51 [49,53]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	23 [9,38]	90 [84,96]	29 [12,45]	55 [46,65]	1 [0,2]	100 [100,100]	1 [0,4]	50 [50,51]
FLARE23 Top 1	3 [0,11]	100 [100,100]	6 [0,20]	52 [50,56]	5 [3,8]	100 [100,100]	10 [5,14]	52 [51,54]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]	15 [10,19]	100 [100,100]	26 [19,32]	57 [55,60]
FLARE24 Top 1	3 [0,12]	100 [100,100]	6 [0,21]	51 [48,55]	4 [2,6]	100 [100,100]	8 [3,12]	52 [51,53]
FLARE25 Baseline	16 [4,30]	93 [88,97]	23 [6,38]	54 [47,61]	2 [1,5]	100 [100,100]	5 [2,9]	51 [50,52]
segmentation from reports (ours)								
R-Super	42 [24,59]	92 [87,97]	49 [30,64]	72 [59,82]	88 [84,92]	88 [81,93]	91 [88,93]	93 [90,96]

AI	spleen small tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	64 [56,71]	42 [33,50]	62 [55,68]	53 [47,60]	62 [58,67]	52 [47,57]	55 [50,59]	57 [53,61]
Merlin	47 [39,56]	58 [48,67]	53 [46,60]	53 [46,60]	42 [38,46]	74 [68,80]	46 [41,51]	58 [52,63]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,51]	0 [0,1]	100 [100,100]	0 [0,1]	50 [50,50]
MedGemma	8 [4,13]	95 [90,98]	15 [8,22]	52 [48,55]	2 [1,3]	99 [98,100]	4 [2,5]	50 [50,51]
Lingshu	25 [18,32]	77 [69,84]	35 [27,43]	51 [46,56]	4 [3,4]	97 [96,98]	5 [4,6]	50 [49,51]
segmentation from masks								
ULS	14 [9,20]	89 [84,95]	24 [16,32]	52 [48,56]	15 [12,19]	91 [89,93]	22 [18,26]	53 [50,55]
FLARE23 Top 1	15 [10,21]	100 [100,100]	26 [18,35]	58 [55,61]	6 [4,8]	100 [99,100]	10 [7,13]	53 [52,54]
FLARE23 Top 2	31 [23,37]	97 [94,100]	46 [38,54]	64 [60,68]	14 [11,16]	99 [98,99]	22 [19,25]	56 [55,57]
FLARE24 Top 1	24 [17,30]	97 [94,100]	38 [29,46]	60 [57,64]	8 [6,10]	99 [99,100]	13 [10,16]	53 [52,54]
FLARE25 Baseline	18 [12,24]	99 [97,100]	30 [21,38]	58 [55,62]	11 [8,14]	97 [95,98]	18 [14,22]	54 [52,56]
segmentation from reports (ours)								
R-Super	77 [70,83]	92 [87,97]	84 [79,89]	86 [80,90]	69 [65,73]	84 [81,87]	71 [67,75]	81 [78,84]

Table D25: **Small tumor detection, external validation (Merlin, Stanford, USA): trained with no mask, R-Super accurately detects small tumors (diameter ≤ 2 cm).** This table evaluates tumor detection (malignant + benign vs no-tumor), evaluating only on small tumors. R-Super was trained with the UCSF dataset (no tumor mask), and it was evaluated on the Merlin test set (Stanford hospital). No data from the Stanford hospital was used in training, making this an external (out-of-distribution) test set. See dataset details in Tab. 1, A1. We compare R-Super to prominent public AI models: MedGemma, Merlin, MedImageInsight, Lingshu, ULS, and winners of the FLARE cancer segmentation challenge. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best, using DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples. n/a: we skipped organs with less than 10 positive scans in the dataset.

AI	bladder small tumor				esophagus small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	83 [71,94]	51 [41,60]	52 [41,62]	68 [57,77]	n/a	n/a	n/a	n/a
Merlin	63 [50,78]	84 [77,90]	61 [48,72]	76 [66,85]	n/a	n/a	n/a	n/a
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]	n/a	n/a	n/a	n/a
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	n/a	n/a	n/a	n/a
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	49 [48,50]	n/a	n/a	n/a	n/a
segmentation from masks								
ULS	29 [15,44]	74 [66,82]	29 [15,42]	44 [33,54]	n/a	n/a	n/a	n/a
FLARE23 Top 1	7 [0,16]	100 [100,100]	14 [0,28]	54 [50,58]	n/a	n/a	n/a	n/a
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	n/a	n/a	n/a	n/a
FLARE24 Top 1	2 [0,9]	99 [97,100]	5 [0,15]	51 [49,54]	n/a	n/a	n/a	n/a
FLARE25 Baseline	10 [2,20]	100 [100,100]	18 [4,33]	55 [51,60]	n/a	n/a	n/a	n/a
segmentation from reports (ours)								
R-Super	73 [60,86]	82 [73,88]	65 [53,75]	81 [72,88]	n/a	n/a	n/a	n/a

AI	gallbladder small tumor				uterus small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	27 [14,41]	89 [83,94]	34 [18,49]	57 [46,67]	70 [57,83]	29 [21,37]	41 [31,51]	48 [39,60]
Merlin	63 [48,78]	69 [61,78]	51 [38,63]	68 [58,78]	47 [33,60]	81 [74,88]	48 [35,60]	67 [57,76]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]
Lingshu	7 [0,17]	98 [96,100]	13 [0,28]	53 [49,57]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	24 [12,38]	92 [87,97]	33 [17,48]	58 [51,66]	40 [28,56]	68 [60,77]	37 [26,49]	54 [46,63]
FLARE23 Top 1	5 [0,12]	100 [100,100]	9 [0,22]	52 [49,56]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE23 Top 2	15 [5,25]	98 [96,100]	24 [9,38]	56 [51,61]	4 [0,11]	99 [97,100]	8 [0,20]	52 [49,55]
FLARE24 Top 1	7 [0,17]	100 [100,100]	14 [0,29]	54 [50,58]	4 [0,11]	99 [97,100]	8 [0,19]	52 [49,55]
FLARE25 Baseline	5 [0,12]	100 [100,100]	9 [0,21]	52 [50,56]	30 [16,43]	87 [81,93]	37 [22,49]	58 [51,66]
segmentation from reports (ours)								
R-Super	68 [54,82]	75 [67,83]	58 [45,68]	76 [66,85]	74 [61,87]	85 [78,91]	71 [59,80]	87 [80,93]

AI	prostate small tumor				adrenal small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	n/a	n/a	n/a	n/a	15 [12,18]	88 [82,93]	25 [20,30]	44 [38,50]
Merlin	n/a	n/a	n/a	n/a	55 [50,60]	85 [78,91]	69 [64,73]	71 [67,76]
VLM report generation								
Merlin	n/a	n/a	n/a	n/a	0 [0,1]	100 [100,100]	0 [0,2]	50 [50,50]
MedGemma	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	49 [46,50]
Lingshu	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	n/a	n/a	n/a	n/a	2 [0,3]	100 [100,100]	3 [1,6]	51 [50,51]
FLARE23 Top 1	n/a	n/a	n/a	n/a	7 [5,10]	100 [100,100]	13 [9,17]	54 [52,55]
FLARE23 Top 2	n/a	n/a	n/a	n/a	18 [14,22]	100 [100,100]	30 [24,35]	59 [57,61]
FLARE24 Top 1	n/a	n/a	n/a	n/a	8 [5,10]	100 [100,100]	14 [9,18]	54 [52,55]
FLARE25 Baseline	n/a	n/a	n/a	n/a	5 [3,7]	100 [100,100]	9 [5,13]	52 [51,53]
segmentation from reports (ours)								
R-Super	n/a	n/a	n/a	n/a	87 [83,90]	91 [86,96]	92 [89,94]	94 [92,96]

AI	spleen small tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	92 [88,96]	18 [11,26]	75 [70,79]	54 [47,62]	50 [41,62]	64 [61,67]	40 [31,47]	54 [46,63]
Merlin	52 [45,60]	88 [82,93]	65 [58,72]	70 [64,76]	55 [46,65]	80 [75,86]	46 [41,52]	70 [61,78]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	50 [48,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [46,54]	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]
Lingshu	27 [21,34]	74 [66,82]	38 [30,45]	50 [45,56]	5 [4,7]	96 [95,97]	7 [5,10]	50 [49,51]
segmentation from masks								
ULS	37 [29,44]	78 [71,85]	48 [40,55]	58 [52,63]	22 [16,33]	84 [81,86]	22 [18,26]	52 [48,58]
FLARE23 Top 1	12 [7,17]	99 [97,100]	21 [13,29]	56 [53,58]	4 [3,6]	100 [100,100]	8 [6,11]	52 [51,53]
FLARE23 Top 2	26 [20,33]	96 [92,99]	40 [33,48]	61 [57,65]	9 [7,11]	99 [98,100]	15 [12,18]	54 [53,55]
FLARE24 Top 1	12 [8,18]	100 [100,100]	22 [14,30]	56 [54,59]	5 [3,7]	100 [99,100]	9 [6,12]	52 [51,53]
FLARE25 Baseline	5 [2,8]	100 [100,100]	9 [4,15]	52 [51,54]	8 [5,10]	96 [95,98]	12 [8,15]	52 [50,54]
segmentation from reports (ours)								
R-Super	70 [63,77]	90 [85,95]	80 [74,84]	83 [79,88]	70 [59,80]	83 [81,86]	58 [52,64]	82 [77,86]

Table D26: **Small tumor detection, external validation (Medipol, Turkey): trained with no mask, R-Super accurately detects small tumors (diameter ≤ 2 cm).** This table evaluates tumor detection (malignant + benign vs no-tumor), evaluating only on small tumors. R-Super was trained with the R-Super train dataset (USA), and it was evaluated on the Medipol test set (external validation, Turkey). See dataset details in Tab. 1, A1. We compare R-Super to prominent public AI models: MedGemma, Merlin, MedImageInsight, Lingshu, ULS, and winners of the FLARE cancer segmentation challenge. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best, using DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples. n/a: we skipped organs with less than 10 positive scans in the dataset.

AI	bladder small tumor				esophagus small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	54 [40,67]	79 [72,84]	48 [37,59]	73 [66,80]	7 [0,15]	94 [90,97]	10 [0,22]	56 [46,66]
Merlin	41 [27,55]	74 [68,80]	36 [24,47]	62 [52,70]	75 [60,87]	48 [41,56]	39 [29,49]	65 [56,74]
VLM report generation								
Merlin	4 [0,10]	99 [98,100]	7 [0,17]	52 [50,54]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
MedGemma	0 [0,0]	99 [98,100]	0 [0,0]	50 [49,50]	0 [0,0]	99 [98,100]	0 [0,0]	50 [49,50]
Lingshu	6 [0,13]	98 [96,100]	10 [0,22]	52 [49,56]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	59 [46,72]	72 [66,80]	47 [36,57]	71 [62,79]	59 [44,74]	86 [82,91]	55 [42,67]	74 [64,82]
FLARE23 Top 1	0 [0,0]	99 [98,100]	0 [0,0]	49 [48,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	14 [4,25]	100 [100,100]	24 [8,40]	57 [52,62]
FLARE24 Top 1	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	2 [0,8]	100 [100,100]	4 [0,14]	51 [50,54]
FLARE25 Baseline	4 [0,9]	99 [98,100]	7 [0,16]	52 [50,54]	2 [0,8]	98 [96,100]	4 [0,13]	50 [48,53]
segmentation from reports (ours)								
R-Super	91 [82,98]	65 [58,71]	59 [49,67]	80 [74,85]	86 [75,96]	89 [84,93]	74 [63,84]	93 [89,97]

AI	gallbladder small tumor				uterus small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	48 [37,59]	82 [76,87]	50 [40,60]	69 [62,76]	100 [100,100]	24 [18,30]	15 [8,23]	47 [37,58]
Merlin	51 [40,61]	69 [63,76]	45 [36,54]	62 [55,69]	42 [14,70]	82 [76,87]	20 [5,36]	69 [55,84]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	49 [48,50]	33 [8,62]	87 [82,92]	20 [5,39]	60 [47,76]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	51 [49,53]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]
segmentation from masks								
ULS	43 [32,54]	82 [76,87]	46 [36,55]	63 [57,69]	50 [21,80]	74 [68,80]	18 [7,31]	66 [50,83]
FLARE23 Top 1	1 [0,4]	98 [96,100]	2 [0,8]	50 [48,52]	0 [0,0]	99 [98,100]	0 [0,0]	49 [48,50]
FLARE23 Top 2	20 [11,30]	97 [94,99]	32 [19,44]	59 [54,64]	0 [0,0]	100 [100,100]	0 [0,0]	50 [49,50]
FLARE24 Top 1	1 [0,4]	99 [98,100]	3 [0,8]	50 [49,52]	0 [0,0]	97 [94,99]	0 [0,0]	49 [47,50]
FLARE25 Baseline	1 [0,4]	98 [96,100]	2 [0,8]	49 [48,51]	0 [0,0]	94 [90,97]	0 [0,0]	47 [45,49]
segmentation from reports (ours)								
R-Super	56 [45,67]	68 [62,75]	49 [39,57]	68 [61,76]	100 [100,100]	76 [70,82]	36 [20,50]	83 [76,88]

AI	prostate small tumor				adrenal small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	n/a	n/a	n/a	n/a	67 [58,76]	49 [42,56]	54 [47,61]	59 [52,66]
Merlin	n/a	n/a	n/a	n/a	57 [48,66]	80 [74,85]	60 [52,68]	72 [65,78]
VLM report generation								
Merlin	n/a	n/a	n/a	n/a	0 [0,0]	99 [97,100]	0 [0,0]	49 [48,50]
MedGemma	n/a	n/a	n/a	n/a	4 [1,8]	98 [95,100]	8 [2,14]	51 [49,53]
Lingshu	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	n/a	n/a	n/a	n/a	12 [6,18]	97 [94,99]	21 [12,30]	55 [51,58]
FLARE23 Top 1	n/a	n/a	n/a	n/a	2 [0,5]	100 [100,100]	4 [0,9]	51 [50,52]
FLARE23 Top 2	n/a	n/a	n/a	n/a	4 [1,7]	100 [100,100]	7 [2,14]	52 [50,54]
FLARE24 Top 1	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE25 Baseline	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from reports (ours)								
R-Super	n/a	n/a	n/a	n/a	94 [89,98]	91 [87,95]	90 [86,94]	96 [93,98]

AI	spleen small tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	76 [69,84]	37 [30,45]	59 [52,65]	53 [46,59]	59 [55,63]	61 [57,64]	40 [36,44]	60 [55,64]
Merlin	77 [70,84]	65 [58,72]	69 [63,75]	71 [65,77]	57 [50,63]	70 [64,75]	45 [40,50]	67 [62,72]
VLM report generation								
Merlin	1 [0,2]	99 [97,100]	1 [0,5]	50 [49,51]	6 [2,11]	97 [96,98]	5 [2,8]	52 [50,54]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [47,52]	1 [0,1]	99 [99,100]	1 [0,2]	50 [49,51]
Lingshu	0 [0,0]	100 [100,100]	0 [0,0]	51 [46,56]	1 [0,2]	100 [99,100]	2 [0,4]	50 [49,51]
segmentation from masks								
ULS	58 [49,67]	51 [43,58]	52 [45,59]	53 [47,59]	47 [41,53]	77 [74,80]	40 [35,44]	64 [60,67]
FLARE23 Top 1	27 [20,35]	97 [94,99]	42 [32,50]	62 [58,66]	5 [4,6]	99 [98,100]	8 [6,10]	52 [51,53]
FLARE23 Top 2	45 [37,54]	90 [86,94]	57 [49,65]	66 [62,72]	14 [11,17]	98 [97,99]	20 [16,24]	56 [54,57]
FLARE24 Top 1	18 [12,24]	95 [92,98]	29 [20,36]	56 [53,60]	4 [2,5]	99 [98,99]	6 [4,8]	51 [50,52]
FLARE25 Baseline	11 [6,16]	96 [92,98]	19 [11,27]	53 [50,56]	3 [2,5]	98 [97,98]	6 [3,8]	50 [49,51]
segmentation from reports (ours)								
R-Super	71 [63,78]	89 [84,94]	77 [70,82]	85 [81,89]	83 [80,86]	80 [77,82]	64 [60,68]	84 [82,86]

Table D27: **Small tumor detection, external validation (Basel, Switzerland): trained with no mask, R-Super accurately detects small tumors (diameter ≤ 2 cm).** This table evaluates tumor detection (malignant + benign vs no-tumor), evaluating only on small tumors. R-Super was trained with the R-Super train dataset (USA), and it was evaluated on the Basel test set (external validation, Switzerland). See dataset details in Tab. 1, A1. We compare R-Super to prominent public AI models: MedGemma, Merlin, MedImageInsight, Lingshu, ULS, and winners of the FLARE cancer segmentation challenge. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best, using DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals, calculated via bootstrapping, with 1,000 bootstrap samples. n/a: we skipped organs with less than 10 positive scans in the dataset.

AI	bladder small tumor				esophagus small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	76 [62,88]	78 [66,89]	73 [60,82]	83 [73,91]	n/a	n/a	n/a	n/a
Merlin	11 [2,21]	93 [85,98]	18 [4,32]	43 [32,55]	n/a	n/a	n/a	n/a
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	49 [47,50]	n/a	n/a	n/a	n/a
MedGemma	5 [0,14]	100 [100,100]	10 [0,24]	53 [50,57]	n/a	n/a	n/a	n/a
Lingshu	14 [3,26]	93 [85,98]	22 [5,38]	53 [46,60]	n/a	n/a	n/a	n/a
segmentation from masks								
ULS	84 [70,95]	65 [52,78]	71 [58,81]	76 [65,85]	n/a	n/a	n/a	n/a
FLARE23 Top 1	5 [0,14]	100 [100,100]	10 [0,24]	53 [50,57]	n/a	n/a	n/a	n/a
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	n/a	n/a	n/a	n/a
FLARE24 Top 1	0 [0,0]	100 [100,100]	0 [0,0]	49 [47,50]	n/a	n/a	n/a	n/a
FLARE25 Baseline	3 [0,9]	98 [94,100]	5 [0,16]	50 [48,54]	n/a	n/a	n/a	n/a
segmentation from reports (ours)								
R-Super	51 [36,68]	92 [84,98]	63 [49,77]	83 [74,91]	n/a	n/a	n/a	n/a

AI	gallbladder small tumor				uterus small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	n/a	n/a	n/a	n/a	83 [60,100]	32 [20,45]	34 [19,49]	41 [26,58]
Merlin	n/a	n/a	n/a	n/a	17 [0,40]	87 [78,95]	19 [0,40]	51 [34,70]
VLM report generation								
Merlin	n/a	n/a	n/a	n/a	17 [0,42]	93 [85,98]	22 [0,47]	55 [45,67]
MedGemma	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Lingshu	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	n/a	n/a	n/a	n/a	42 [14,70]	94 [88,100]	50 [18,73]	72 [54,88]
FLARE23 Top 1	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE23 Top 2	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE24 Top 1	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
FLARE25 Baseline	n/a	n/a	n/a	n/a	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from reports (ours)								
R-Super	n/a	n/a	n/a	n/a	42 [14,71]	100 [100,100]	59 [25,83]	78 [54,97]

AI	prostate small tumor				adrenal small tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	50 [21,76]	70 [57,82]	38 [16,57]	57 [39,75]	16 [10,22]	87 [76,95]	26 [17,34]	40 [31,48]
Merlin	7 [0,23]	100 [100,100]	13 [0,38]	27 [13,44]	12 [7,17]	89 [80,96]	20 [12,28]	39 [30,47]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	48 [45,50]	0 [0,0]	100 [100,100]	0 [0,0]	49 [46,51]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Lingshu	14 [0,36]	96 [90,100]	22 [0,47]	55 [47,66]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
segmentation from masks								
ULS	64 [38,90]	87 [77,96]	60 [36,79]	77 [61,92]	27 [20,35]	96 [91,100]	42 [33,51]	56 [48,62]
FLARE23 Top 1	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	4 [1,7]	98 [94,100]	7 [2,13]	51 [48,53]
FLARE23 Top 2	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	8 [4,13]	100 [100,100]	15 [8,22]	54 [52,56]
FLARE24 Top 1	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	4 [1,7]	100 [100,100]	7 [2,13]	52 [50,54]
FLARE25 Baseline	14 [0,36]	100 [100,100]	25 [0,53]	57 [50,68]	3 [1,6]	100 [100,100]	6 [1,12]	52 [50,53]
segmentation from reports (ours)								
R-Super	69 [42,94]	89 [79,96]	64 [40,83]	82 [66,95]	65 [57,74]	85 [74,94]	76 [70,82]	79 [73,85]

AI	spleen small tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
VLM zero-shot classification								
MedImageInsight	92 [87,97]	17 [7,26]	79 [74,84]	42 [32,52]	63 [55,70]	57 [51,62]	50 [43,57]	52 [46,59]
Merlin	12 [7,19]	91 [82,98]	21 [12,30]	42 [34,51]	12 [7,17]	92 [86,96]	18 [11,26]	40 [32,49]
VLM report generation								
Merlin	0 [0,0]	100 [100,100]	0 [0,0]	49 [46,52]	3 [0,8]	98 [97,100]	4 [0,9]	50 [47,53]
MedGemma	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]	1 [0,3]	100 [100,100]	2 [0,5]	50 [50,51]
Lingshu	18 [11,25]	80 [68,90]	28 [19,37]	49 [42,55]	9 [5,14]	94 [90,97]	14 [8,21]	51 [49,55]
segmentation from masks								
ULS	80 [72,86]	17 [7,27]	73 [66,78]	32 [22,42]	59 [51,67]	72 [67,76]	59 [50,67]	62 [56,68]
FLARE23 Top 1	23 [16,31]	94 [88,100]	37 [26,46]	59 [54,64]	6 [4,9]	98 [97,100]	11 [7,14]	53 [51,54]
FLARE23 Top 2	38 [28,47]	85 [76,94]	52 [42,62]	61 [55,68]	9 [7,11]	97 [95,99]	14 [11,16]	53 [52,55]
FLARE24 Top 1	15 [8,22]	82 [71,92]	24 [15,33]	47 [41,54]	4 [2,5]	96 [94,98]	6 [4,8]	50 [48,51]
FLARE25 Baseline	8 [3,14]	85 [75,94]	14 [6,22]	45 [38,51]	6 [2,10]	97 [94,99]	10 [4,16]	51 [48,53]
segmentation from reports (ours)								
R-Super	60 [50,68]	85 [75,94]	71 [64,78]	75 [67,82]	58 [48,66]	90 [86,94]	67 [57,74]	79 [73,85]

D.7 Comparison to Alternative Training Methods Trained on Our Dataset

After training R-Super with only CT-Report pairs, we used it to help 31 radiologists create tumor segmentation masks for our dataset. With the help of R-Super, radiologists were able to create masks about 5 times faster (Sec. C.2). They created a large number of masks, over 1,400 (Tab. 1). This number surpasses even the largest public pancreatic tumor dataset, PanTS [18]. However, these masks still represent less than 2% of our number of CT-Report pairs, exemplifying the difficulty of mask creation and the importance of learning from radiology reports. In this section, we present detailed results comparing R-Super (trained only on CT-Report pairs) and R-Super† (trained on CT-Report plus CT-Mask pairs) to alternative training methods, which were also trained on the datasets that we built. We show that R-Super more effectively learns from reports and masks, consistently achieving superior performance in tumor detection, malignant tumor detection, and small tumor detection.

D.7.1 Tumor Detection

Table D28: **Tumor detection, internal validation (UCSF, USA): comparison to alternative methods trained on the same dataset.** This table evaluates tumor detection (malignant + benign vs no-tumor). All models were trained on the R-Super train set (UCSF + Stanford), and evaluated on the UCSF test set (internal validation). See dataset details in Tab. 1, A1. All models were evaluated at 90% specificity. We compare multiple training methods: R-Super†, MTL, and CLIP were trained on all available CT-Report plus CT-Mask pairs; segmentation and nnU-Net on CT-Mask pairs only; R-Super and classification on CT-Report only; ModelsGenesis on all CT scans and all CT-Mask pairs. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model (separately for models trained without masks). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% CI (1,000 bootstrap samples).

AI	bladder tumor			esophagus tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	23 [19,28]	37 [31,42]	68 [62,74]	63 [58,68]	76 [72,80]	85 [81,89]
R-Super	51 [45,56]	66 [61,71]	79 [74,83]	80 [76,84]	88 [85,90]	91 [88,94]
many reports, few masks training						
CLIP	34 [29,39]	49 [44,55]	73 [68,78]	73 [68,78]	83 [80,86]	88 [84,91]
MTL	40 [36,46]	56 [51,62]	76 [72,81]	75 [71,80]	84 [82,87]	91 [88,94]
ModelsGenesis	41 [35,46]	56 [51,62]	76 [71,80]	70 [66,74]	81 [78,84]	89 [85,92]
segmentation	42 [37,48]	58 [52,63]	76 [71,80]	79 [74,83]	87 [84,90]	87 [83,90]
nnUNet	36 [31,41]	52 [46,57]	74 [69,78]	75 [70,79]	84 [81,87]	87 [83,89]
R-Super†	64 [59,69]	76 [73,80]	82 [78,86]	82 [79,86]	89 [86,91]	90 [88,93]

AI	gallbladder tumor			uterus tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	4 [2,7]	8 [4,12]	50 [44,57]	50 [44,55]	65 [60,70]	79 [74,83]
R-Super	51 [46,57]	66 [60,71]	78 [73,82]	78 [73,83]	86 [83,89]	89 [86,92]
many reports, few masks training						
CLIP	48 [43,55]	64 [58,69]	77 [72,82]	64 [59,70]	77 [72,81]	86 [81,90]
MTL	60 [54,66]	73 [68,77]	80 [76,84]	70 [65,75]	81 [77,84]	89 [86,92]
ModelsGenesis	52 [45,57]	66 [60,71]	79 [75,84]	67 [62,72]	78 [74,82]	87 [83,90]
segmentation	59 [53,64]	72 [67,76]	80 [75,84]	68 [62,73]	79 [75,83]	88 [84,91]
nnUNet	43 [36,48]	58 [52,63]	68 [62,72]	61 [55,66]	74 [70,78]	83 [79,87]
R-Super†	66 [59,71]	77 [72,81]	84 [80,88]	80 [75,84]	87 [84,90]	88 [84,91]

AI	prostate tumor			adrenal tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	19 [14,25]	31 [24,39]	74 [68,80]	9 [7,11]	16 [13,20]	47 [41,52]
R-Super	70 [63,77]	80 [74,84]	85 [80,90]	84 [82,87]	91 [89,92]	93 [90,95]
many reports, few masks training						
CLIP	57 [50,64]	70 [64,76]	81 [76,86]	71 [68,74]	82 [80,85]	89 [86,92]
MTL	61 [54,68]	73 [67,78]	85 [81,89]	77 [74,80]	86 [84,88]	91 [88,93]
ModelsGenesis	58 [50,65]	70 [64,76]	83 [78,88]	73 [69,77]	84 [81,86]	90 [87,92]
segmentation	61 [54,68]	73 [67,78]	83 [79,88]	81 [78,84]	89 [87,90]	92 [89,94]
nnUNet	53 [46,60]	67 [60,73]	79 [74,84]	62 [59,66]	76 [73,79]	79 [76,82]
R-Super†	66 [59,73]	77 [71,82]	87 [82,91]	83 [81,86]	90 [88,92]	91 [89,93]

AI	spleen tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	11 [8,14]	20 [15,25]	50 [44,56]	26 [24,28]	36 [34,38]	65 [62,68]
R-Super	77 [73,81]	86 [83,88]	88 [85,91]	70 [68,72]	80 [79,82]	86 [84,88]
many reports, few masks training						
CLIP	36 [32,41]	53 [48,57]	74 [69,78]	55 [53,57]	68 [66,70]	81 [79,83]
MTL	53 [49,58]	68 [65,73]	80 [76,83]	62 [61,64]	75 [73,76]	85 [83,86]
ModelsGenesis	33 [29,38]	49 [44,54]	76 [72,81]	56 [54,58]	69 [67,71]	83 [81,85]
segmentation	46 [41,50]	62 [57,66]	80 [75,84]	62 [60,64]	74 [72,76]	84 [82,85]
nnUNet	37 [33,42]	54 [49,58]	73 [67,77]	52 [50,55]	66 [64,68]	77 [76,79]
R-Super†	77 [73,81]	86 [83,88]	87 [84,90]	74 [72,76]	83 [82,84]	87 [86,88]

Table D29: **Tumor detection, external validation (Stanford, USA): comparison to alternative methods trained on the same dataset.** This table evaluates tumor detection (malignant + benign vs no-tumor). All models were trained on the UCSF train set (USA), and evaluated on the Stanford test set (external validation, USA). Dataset details in Tab. 1 and A1. All models were evaluated at 90% specificity. We compare multiple training methods: R-Super†, MTL, and CLIP were trained on all available CT-Report plus CT-Mask pairs; segmentation and nnU-Net on CT-Mask pairs only; R-Super and classification on CT-Report only; ModelsGenesis on all CT scans and all CT-Mask pairs. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model (separately for models trained without masks). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% CI (1,000 bootstrap samples).

AI	bladder tumor			esophagus tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	4 [2,7]	7 [3,12]	45 [39,52]	29 [17,44]	38 [22,52]	69 [60,78]
R-Super	61 [54,68]	73 [68,79]	82 [77,87]	44 [29,59]	51 [36,65]	79 [70,87]
many reports, few masks training						
CLIP	35 [29,41]	50 [43,56]	76 [71,81]	76 [61,88]	75 [63,85]	90 [83,96]
MTL	71 [64,77]	81 [76,85]	83 [79,86]	76 [62,88]	76 [65,86]	85 [78,92]
ModelsGenesis	39 [32,45]	54 [47,60]	74 [69,79]	78 [64,90]	77 [65,86]	88 [80,95]
segmentation	45 [39,52]	60 [53,66]	75 [70,80]	68 [54,82]	71 [59,81]	86 [78,93]
nnUNet	28 [22,35]	42 [35,49]	65 [59,70]	50 [34,66]	62 [46,75]	74 [66,82]
R-Super†	64 [58,71]	76 [71,81]	84 [79,88]	78 [65,91]	76 [64,86]	90 [83,95]

AI	gallbladder tumor			uterus tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	12 [8,17]	20 [13,27]	50 [43,56]	17 [13,21]	29 [23,34]	71 [64,77]
R-Super	54 [48,61]	68 [62,74]	80 [75,85]	68 [62,72]	79 [75,82]	90 [87,93]
many reports, few masks training						
CLIP	58 [51,64]	71 [65,76]	76 [71,80]	43 [38,48]	59 [54,63]	81 [76,86]
MTL	41 [34,48]	56 [49,62]	70 [65,74]	55 [50,61]	70 [66,74]	85 [81,89]
ModelsGenesis	58 [52,65]	71 [66,77]	78 [74,82]	44 [39,49]	60 [55,65]	82 [78,87]
segmentation	60 [53,66]	73 [67,78]	78 [74,81]	41 [37,47]	57 [52,63]	83 [79,88]
nnUNet	0 [0,0]	0 [0,0]	50 [50,50]	58 [53,63]	72 [68,76]	80 [75,84]
R-Super†	71 [65,77]	80 [76,85]	87 [83,91]	69 [64,74]	80 [77,84]	89 [84,92]

AI	prostate tumor			adrenal tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	36 [28,46]	49 [39,58]	81 [74,86]	15 [12,17]	26 [22,29]	55 [49,60]
R-Super	76 [68,84]	82 [75,87]	89 [84,92]	84 [82,87]	91 [89,92]	93 [90,95]
many reports, few masks training						
CLIP	41 [32,50]	54 [44,63]	80 [74,86]	69 [66,72]	81 [79,83]	82 [80,84]
MTL	72 [64,80]	79 [72,85]	86 [81,91]	80 [77,82]	88 [86,90]	88 [86,90]
ModelsGenesis	60 [50,69]	70 [62,77]	83 [78,88]	69 [65,72]	81 [79,84]	83 [81,85]
segmentation	60 [51,70]	71 [63,79]	85 [79,90]	74 [71,76]	84 [82,86]	86 [84,87]
nnUNet	57 [48,67]	68 [60,75]	83 [77,88]	53 [50,57]	69 [66,72]	75 [72,78]
R-Super†	59 [50,68]	70 [62,77]	90 [86,93]	81 [78,84]	89 [87,90]	90 [88,93]

AI	spleen tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	10 [8,13]	18 [14,23]	51 [45,56]	18 [15,20]	27 [24,30]	60 [58,63]
R-Super	68 [64,72]	80 [77,83]	81 [77,84]	65 [62,68]	75 [72,78]	85 [83,86]
many reports, few masks training						
CLIP	57 [53,62]	72 [68,75]	76 [73,79]	54 [51,57]	66 [63,69]	80 [78,82]
MTL	64 [59,68]	77 [74,80]	80 [77,83]	66 [63,68]	75 [73,78]	82 [81,84]
ModelsGenesis	65 [60,69]	78 [74,81]	80 [77,83]	59 [56,62]	70 [68,73]	81 [79,83]
segmentation	60 [56,64]	74 [71,77]	78 [75,80]	58 [55,62]	70 [67,73]	81 [80,83]
nnUNet	47 [43,52]	63 [59,67]	74 [70,78]	42 [39,45]	54 [51,57]	72 [70,74]
R-Super†	74 [70,78]	84 [81,86]	83 [80,86]	71 [68,74]	79 [77,82]	88 [86,89]

Table D30: **Tumor detection, external validation (Medipol, Turkey): comparison to alternative methods trained on the same dataset.** This table evaluates tumor detection (malignant + benign vs no-tumor). All models were trained on the R-Super train set (USA), and evaluated on the Medipol test set (external validation, Turkey). See dataset details in Tab. 1, A1. All models were evaluated at 90% specificity. We compare multiple training methods: R-Super†, MTL, and CLIP were trained on all available CT-Report plus CT-Mask pairs; segmentation and nnU-Net on CT-Mask pairs only; R-Super and classification on CT-Report only; ModelsGenesis on all CT scans and all CT-Mask pairs. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model (separately for models trained without masks). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% CI (1,000 bootstrap samples).

AI	bladder tumor			esophagus tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	16 [10,22]	25 [17,33]	68 [62,73]	31 [25,37]	44 [37,52]	74 [68,78]
R-Super	59 [51,66]	69 [62,75]	84 [80,88]	72 [66,79]	80 [75,84]	89 [86,92]
many reports, few masks training						
CLIP	27 [20,34]	39 [30,46]	74 [68,79]	67 [60,74]	76 [70,81]	86 [82,90]
MTL	59 [52,67]	69 [62,75]	80 [76,85]	71 [64,77]	79 [74,83]	88 [84,92]
ModelsGenesis	36 [28,44]	48 [40,57]	72 [67,78]	65 [58,72]	75 [69,79]	87 [83,90]
segmentation	40 [32,48]	52 [44,61]	79 [74,84]	67 [60,74]	76 [71,82]	88 [84,91]
nnUNet	46 [37,53]	58 [50,65]	73 [68,78]	58 [51,65]	69 [63,75]	79 [74,83]
R-Super†	61 [53,69]	71 [64,77]	87 [83,90]	65 [57,72]	74 [69,80]	89 [86,92]

AI	gallbladder tumor			uterus tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	3 [1,6]	5 [1,10]	34 [28,39]	49 [40,58]	60 [52,68]	79 [74,84]
R-Super	34 [27,41]	47 [39,55]	68 [63,74]	57 [48,66]	67 [60,74]	88 [84,92]
many reports, few masks training						
CLIP	34 [28,42]	47 [40,55]	76 [72,81]	58 [49,67]	68 [60,75]	85 [81,89]
MTL	32 [26,40]	45 [38,53]	71 [66,76]	65 [57,73]	73 [66,78]	89 [86,93]
ModelsGenesis	22 [16,28]	33 [25,41]	71 [65,76]	59 [51,67]	68 [62,75]	85 [81,89]
segmentation	38 [31,46]	52 [44,59]	72 [66,77]	67 [59,75]	75 [68,81]	88 [85,92]
nnUNet	14 [8,19]	22 [14,30]	60 [54,65]	65 [58,74]	73 [67,80]	82 [78,87]
R-Super†	57 [49,64]	68 [62,74]	77 [72,82]	59 [50,67]	70 [62,76]	90 [87,93]

AI	prostate tumor			adrenal tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	34 [19,50]	38 [22,51]	84 [78,90]	14 [10,18]	23 [18,29]	58 [53,63]
R-Super	83 [70,95]	74 [62,85]	93 [86,98]	88 [84,91]	91 [89,94]	94 [91,96]
many reports, few masks training						
CLIP	89 [77,97]	76 [65,85]	93 [86,99]	75 [71,80]	84 [81,87]	92 [89,94]
MTL	80 [67,92]	70 [57,81]	94 [89,98]	86 [82,90]	90 [88,93]	93 [90,95]
ModelsGenesis	94 [85,100]	76 [65,87]	96 [92,99]	81 [77,85]	88 [84,90]	92 [89,95]
segmentation	89 [78,97]	75 [63,84]	92 [84,98]	81 [76,85]	87 [84,90]	93 [90,95]
nnUNet	73 [56,88]	67 [51,78]	83 [73,91]	69 [64,74]	79 [76,83]	82 [78,85]
R-Super†	89 [77,98]	75 [64,85]	96 [92,99]	83 [79,86]	88 [86,91]	90 [88,93]

AI	spleen tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	9 [6,12]	16 [11,20]	51 [46,56]	22 [19,25]	30 [27,33]	64 [61,66]
R-Super	59 [54,64]	72 [68,77]	81 [77,84]	65 [62,68]	72 [69,74]	85 [84,87]
many reports, few masks training						
CLIP	20 [16,24]	32 [27,38]	71 [66,75]	53 [50,56]	60 [57,63]	83 [80,84]
MTL	32 [27,36]	46 [41,51]	75 [72,79]	61 [58,64]	68 [65,70]	84 [83,86]
ModelsGenesis	11 [8,15]	20 [15,24]	71 [66,76]	53 [50,55]	58 [55,61]	82 [80,84]
segmentation	17 [13,22]	28 [23,34]	75 [70,80]	57 [54,60]	64 [61,66]	84 [82,86]
nnUNet	36 [30,41]	51 [45,56]	70 [65,74]	51 [48,55]	60 [57,63]	76 [73,78]
R-Super†	55 [50,60]	68 [64,73]	80 [76,84]	67 [64,70]	74 [71,76]	87 [86,89]

Table D31: **Tumor detection, external validation (Basel, Switzerland): comparison to alternative methods trained on the same dataset.** This table evaluates tumor detection (malignant + benign vs no-tumor). All models were trained on the R-Super train set (USA), and evaluated on the Basel test set (external validation, Switzerland). See dataset details in Tab. 1, A1. All models were evaluated at 90% specificity. We compare multiple training methods: R-Super†, MTL, and CLIP were trained on all available CT-Report plus CT-Mask pairs; segmentation and nnU-Net on CT-Mask pairs only; R-Super and classification on CT-Report only; ModelsGenesis on all CT scans and all CT-Mask pairs. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model (separately for models trained without masks). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% CI (1,000 bootstrap samples).

AI	bladder tumor			esophagus tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	24 [19,29]	38 [32,45]	81 [72,88]	61 [50,71]	73 [64,81]	85 [78,92]
R-Super	72 [66,77]	83 [79,86]	89 [84,92]	70 [60,79]	80 [72,86]	90 [84,94]
many reports, few masks training						
CLIP	83 [78,87]	90 [87,92]	91 [87,94]	72 [62,81]	81 [73,87]	85 [77,91]
MTL	82 [78,86]	89 [86,92]	91 [87,93]	78 [69,86]	85 [78,90]	86 [79,92]
ModelsGenesis	71 [66,76]	82 [78,86]	90 [86,93]	72 [62,81]	81 [74,87]	84 [76,90]
segmentation	73 [68,78]	84 [80,87]	88 [84,92]	77 [67,86]	84 [77,90]	87 [80,92]
nnUNet	80 [75,84]	88 [85,91]	88 [84,92]	75 [66,84]	83 [76,89]	87 [80,93]
R-Super†	86 [82,90]	92 [89,94]	92 [88,94]	76 [66,84]	84 [76,89]	92 [87,96]

AI	gallbladder tumor			uterus tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	1 [0,4]	2 [0,8]	36 [26,46]	65 [60,71]	78 [74,82]	86 [81,90]
R-Super	26 [17,36]	40 [27,51]	64 [55,75]	70 [63,75]	81 [77,85]	83 [78,87]
many reports, few masks training						
CLIP	49 [37,61]	63 [52,72]	76 [67,83]	86 [82,90]	91 [89,94]	92 [89,94]
MTL	42 [31,54]	58 [46,68]	75 [65,83]	85 [80,89]	91 [88,94]	92 [89,94]
ModelsGenesis	41 [30,52]	55 [44,66]	72 [63,80]	85 [80,89]	91 [88,93]	91 [88,94]
segmentation	36 [25,47]	50 [38,61]	71 [61,79]	86 [82,90]	92 [89,94]	92 [90,94]
nnUNet	41 [30,52]	56 [44,66]	69 [60,77]	78 [74,84]	87 [84,90]	88 [84,91]
R-Super†	59 [48,70]	71 [61,80]	84 [76,90]	85 [80,89]	91 [88,94]	90 [86,93]

AI	prostate tumor			adrenal tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	61 [54,68]	75 [68,80]	86 [79,91]	9 [8,11]	17 [15,20]	49 [42,58]
R-Super	70 [64,77]	81 [76,86]	88 [83,93]	32 [29,34]	48 [45,51]	83 [78,88]
many reports, few masks training						
CLIP	52 [44,60]	67 [60,74]	84 [78,89]	77 [75,80]	87 [86,89]	89 [87,92]
MTL	65 [58,72]	78 [72,83]	84 [78,88]	75 [73,78]	86 [84,87]	87 [83,90]
ModelsGenesis	54 [47,62]	69 [62,75]	84 [77,89]	77 [75,79]	87 [85,88]	89 [85,93]
segmentation	57 [49,65]	71 [64,77]	84 [78,89]	82 [80,85]	90 [89,92]	90 [86,93]
nnUNet	55 [48,63]	70 [63,76]	77 [70,82]	68 [65,70]	80 [79,82]	82 [78,86]
R-Super†	66 [59,73]	78 [73,83]	86 [81,91]	78 [76,80]	88 [86,89]	87 [83,90]

AI	spleen tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	5 [3,6]	9 [6,12]	46 [37,56]	32 [30,35]	42 [40,44]	67 [64,70]
R-Super	58 [54,61]	73 [70,76]	82 [77,86]	57 [54,59]	69 [67,72]	83 [80,85]
many reports, few masks training						
CLIP	22 [19,26]	36 [32,40]	64 [56,72]	63 [60,66]	74 [71,76]	83 [81,85]
MTL	9 [7,12]	17 [13,20]	57 [49,66]	62 [60,65]	72 [70,74]	82 [79,84]
ModelsGenesis	17 [14,19]	28 [24,32]	61 [53,68]	60 [57,62]	70 [68,73]	81 [79,84]
segmentation	15 [12,18]	26 [22,30]	62 [54,71]	61 [58,64]	71 [68,73]	82 [80,84]
nnUNet	28 [25,31]	43 [39,47]	62 [54,69]	61 [58,63]	72 [70,75]	79 [76,81]
R-Super†	55 [51,58]	70 [67,74]	82 [77,87]	72 [69,75]	82 [80,84]	88 [86,89]

D.7.2 Malignant Tumor Detection

Table D32: **Malignant tumor detection, internal validation (UCSF, USA): comparison to alternative methods trained on the same dataset.** This table evaluates tumor detection stratified for malignant tumors (malignant vs no-tumor). All models were trained on the R-Super train set (UCSF + Stanford), and evaluated on the UCSF test set (internal validation). See dataset details in Tab. 1, A1. All models were evaluated at 90% specificity. We compare multiple training methods: R-Super†, MTL, and CLIP were trained on all available CT-Report plus CT-Mask pairs; segmentation and nnU-Net on CT-Mask pairs only; R-Super and classification on CT-Report only; ModelsGenesis on all CT scans and all CT-Mask pairs. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model (separately for models trained without masks). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% CI (1,000 bootstrap samples).

AI	bladder malignant tumor			esophagus malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	25 [19,30]	38 [30,45]	69 [62,75]	67 [61,74]	78 [73,83]	86 [82,90]
R-Super	60 [53,67]	73 [67,78]	82 [77,86]	83 [78,87]	88 [85,91]	92 [89,95]
many reports, few masks training						
CLIP	39 [33,46]	54 [47,61]	77 [71,81]	77 [71,82]	84 [80,88]	90 [86,93]
MTL	47 [40,54]	61 [54,68]	80 [74,85]	77 [72,83]	85 [81,89]	92 [89,95]
ModelsGenesis	47 [40,53]	61 [54,67]	79 [74,84]	70 [64,76]	80 [76,84]	91 [87,94]
segmentation	46 [39,52]	60 [53,66]	79 [74,84]	82 [77,86]	88 [84,91]	88 [84,92]
nnUNet	45 [37,52]	60 [52,66]	79 [74,84]	75 [69,80]	83 [79,87]	87 [83,90]
R-Super†	75 [69,81]	83 [79,87]	87 [83,91]	86 [82,90]	90 [87,93]	94 [91,96]

AI	gallbladder malignant tumor			uterus malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	0 [0,0]	0 [0,0]	51 [43,59]	62 [51,72]	71 [61,78]	84 [77,90]
R-Super	90 [82,96]	88 [82,93]	95 [91,99]	83 [75,91]	85 [78,91]	91 [86,95]
many reports, few masks training						
CLIP	86 [78,93]	86 [79,91]	94 [90,97]	76 [67,85]	80 [74,87]	90 [86,95]
MTL	96 [91,100]	92 [86,96]	97 [95,99]	82 [74,90]	84 [77,90]	92 [87,96]
ModelsGenesis	95 [89,100]	90 [84,95]	97 [95,99]	76 [67,85]	80 [73,87]	90 [85,94]
segmentation	96 [91,100]	92 [86,96]	96 [93,98]	75 [66,84]	80 [73,86]	91 [86,95]
nnUNet	70 [59,80]	77 [69,84]	85 [78,90]	68 [56,78]	75 [67,82]	87 [82,92]
R-Super†	94 [88,98]	90 [85,95]	98 [97,99]	82 [74,91]	84 [78,90]	92 [87,96]

AI	prostate malignant tumor			adrenal malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	24 [15,32]	35 [24,46]	78 [72,85]	14 [8,19]	22 [13,30]	60 [53,67]
R-Super	70 [60,79]	76 [69,83]	88 [82,93]	93 [89,97]	92 [89,96]	96 [94,98]
many reports, few masks training						
CLIP	53 [42,63]	64 [55,73]	79 [73,85]	71 [64,79]	80 [75,86]	88 [84,92]
MTL	62 [51,72]	71 [62,78]	86 [80,91]	83 [77,89]	87 [82,91]	91 [87,95]
ModelsGenesis	52 [41,62]	62 [52,71]	82 [75,88]	79 [72,86]	85 [80,89]	90 [87,94]
segmentation	52 [42,63]	63 [53,72]	81 [75,87]	87 [82,93]	89 [85,93]	93 [89,96]
nnUNet	55 [44,65]	66 [56,74]	79 [72,85]	69 [61,77]	78 [72,83]	80 [74,85]
R-Super†	74 [66,82]	80 [73,86]	90 [85,94]	95 [91,98]	93 [90,96]	96 [94,99]

AI	spleen malignant tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	3 [0,10]	4 [0,14]	42 [32,54]	28 [25,30]	36 [32,38]	67 [64,71]
R-Super	91 [81,100]	82 [71,90]	95 [90,98]	81 [78,84]	84 [81,86]	91 [89,93]
many reports, few masks training						
CLIP	56 [39,73]	60 [45,74]	80 [70,89]	65 [62,69]	73 [69,76]	85 [83,88]
MTL	74 [58,89]	71 [58,83]	85 [76,93]	74 [71,78]	79 [76,82]	89 [87,91]
ModelsGenesis	52 [35,68]	56 [41,69]	81 [72,90]	67 [63,71]	74 [70,76]	87 [85,89]
segmentation	71 [56,86]	70 [56,81]	86 [78,93]	73 [69,76]	77 [74,80]	88 [86,90]
nnUNet	38 [22,56]	46 [29,62]	71 [59,82]	60 [56,64]	69 [66,73]	81 [79,84]
R-Super†	85 [72,96]	78 [67,87]	93 [86,98]	84 [82,87]	86 [83,88]	93 [91,94]

Table D33: **Malignant tumor detection, external validation (Stanford, USA): comparison to alternative methods trained on the same dataset.** This table evaluates tumor detection stratified for malignant tumors (malignant vs no-tumor). All models were trained on the UCSF train set (USA), and evaluated on the Stanford test set (external validation, USA). See dataset details in Tab. 1, A1. All models were evaluated at 90% specificity. We compare multiple training methods: R-Super†, MTL, and CLIP were trained on all available CT-Report plus CT-Mask pairs; segmentation and nnU-Net on CT-Mask pairs only; R-Super and classification on CT-Report only; ModelsGenesis on all CT scans and all CT-Mask pairs. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model (separately for models trained without masks). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% CI (1,000 bootstrap samples).

AI	bladder malignant tumor			esophagus malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	9 [2,18]	14 [3,26]	47 [38,56]	27 [7,50]	27 [7,46]	70 [56,84]
R-Super	73 [61,85]	76 [65,84]	84 [76,91]	60 [33,83]	51 [29,70]	83 [68,94]
many reports, few masks training						
CLIP	54 [41,68]	62 [50,73]	84 [78,90]	87 [67,100]	67 [47,83]	93 [82,100]
MTL	86 [76,94]	85 [77,91]	92 [86,96]	93 [78,100]	74 [54,88]	95 [87,100]
ModelsGenesis	66 [53,78]	71 [60,80]	88 [82,93]	93 [78,100]	72 [53,87]	93 [81,99]
segmentation	78 [67,89]	80 [70,87]	88 [82,94]	93 [76,100]	72 [53,86]	97 [94,99]
nnUNet	47 [34,61]	56 [44,68]	76 [68,84]	87 [69,100]	81 [63,94]	93 [83,100]
R-Super†	86 [75,94]	84 [76,91]	94 [88,98]	93 [78,100]	70 [48,84]	98 [95,100]

AI	gallbladder malignant tumor			uterus malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	22 [10,36]	30 [15,45]	58 [46,68]	12 [4,22]	19 [7,31]	71 [62,79]
R-Super	78 [65,90]	77 [67,86]	92 [86,97]	81 [70,91]	81 [72,88]	95 [92,98]
many reports, few masks training						
CLIP	83 [71,94]	79 [68,88]	90 [83,96]	49 [35,63]	58 [44,69]	80 [72,86]
MTL	58 [43,74]	63 [49,75]	82 [73,89]	75 [64,86]	78 [68,86]	92 [89,96]
ModelsGenesis	85 [73,95]	80 [70,88]	91 [84,97]	49 [36,63]	58 [45,70]	83 [76,89]
segmentation	85 [72,96]	83 [74,90]	92 [85,97]	44 [31,57]	54 [41,66]	82 [75,88]
nnUNet	0 [0,0]	0 [0,0]	50 [50,50]	62 [48,74]	68 [56,78]	82 [75,89]
R-Super†	93 [84,100]	84 [75,92]	96 [91,100]	82 [71,92]	82 [73,89]	92 [87,96]

AI	prostate malignant tumor			adrenal malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	24 [5,44]	27 [6,45]	79 [68,88]	11 [5,17]	17 [8,28]	55 [47,62]
R-Super	90 [76,100]	74 [60,86]	95 [89,99]	97 [93,100]	93 [89,96]	98 [95,100]
many reports, few masks training						
CLIP	62 [41,82]	58 [38,73]	88 [80,94]	90 [84,96]	90 [85,95]	94 [90,97]
MTL	81 [62,96]	69 [53,83]	91 [81,98]	94 [88,98]	93 [88,96]	96 [93,98]
ModelsGenesis	71 [50,90]	64 [44,78]	85 [73,95]	93 [87,98]	93 [89,96]	96 [93,99]
segmentation	67 [47,87]	64 [46,79]	88 [78,97]	92 [86,97]	93 [89,96]	96 [93,98]
nnUNet	67 [46,87]	61 [43,76]	85 [75,94]	76 [67,85]	81 [74,87]	86 [81,92]
R-Super†	86 [68,100]	72 [56,85]	97 [93,100]	97 [93,100]	93 [89,97]	98 [95,99]

AI	spleen malignant tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	9 [3,16]	14 [4,25]	55 [46,64]	16 [11,22]	21 [15,27]	62 [58,66]
R-Super	88 [81,96]	86 [80,92]	92 [88,96]	81 [76,86]	77 [72,82]	91 [89,94]
many reports, few masks training						
CLIP	88 [79,95]	86 [80,92]	93 [88,97]	73 [68,78]	72 [66,76]	89 [86,91]
MTL	88 [81,95]	90 [84,95]	93 [89,97]	82 [78,86]	79 [74,83]	92 [89,94]
ModelsGenesis	91 [84,97]	89 [84,94]	94 [90,98]	78 [73,83]	75 [70,80]	90 [87,92]
segmentation	87 [78,94]	86 [80,92]	92 [87,96]	78 [73,83]	76 [71,80]	91 [88,93]
nnUNet	71 [60,81]	76 [68,84]	86 [80,92]	58 [54,63]	61 [56,64]	80 [77,83]
R-Super†	96 [90,100]	90 [85,95]	96 [93,99]	90 [86,94]	82 [77,86]	96 [94,97]

Table D34: **Malignant tumor detection, external validation (Medipol, Turkey): comparison to alternative methods trained on the same dataset.** This table evaluates tumor detection stratified for malignant tumors (malignant vs no-tumor). All models were trained on the R-Super train set (USA), and evaluated on the Medipol test set (external validation, Turkey). See dataset details in Tab. 1, A1. All models were evaluated at 90% specificity. We compare multiple training methods: R-Super†, MTL, and CLIP were trained on all available CT-Report plus CT-Mask pairs; segmentation and nnU-Net on CT-Mask pairs only; R-Super and classification on CT-Report only; ModelsGenesis on all CT scans and all CT-Mask pairs. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model (separately for models trained without masks). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% CI (1,000 bootstrap samples). n/a: we skipped organs with less than 10 positive scans in the dataset.

AI	bladder malignant tumor			esophagus malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	15 [4,28]	19 [6,33]	72 [64,79]	30 [18,42]	37 [24,49]	74 [65,81]
R-Super	82 [70,94]	73 [61,82]	90 [83,96]	93 [85,98]	83 [74,90]	97 [93,99]
many reports, few masks training						
CLIP	46 [30,62]	49 [33,61]	81 [73,88]	81 [69,91]	76 [66,84]	89 [81,95]
MTL	79 [66,91]	71 [58,80]	89 [81,95]	83 [73,93]	78 [69,86]	93 [87,97]
ModelsGenesis	59 [43,74]	58 [45,71]	86 [80,92]	82 [70,92]	76 [67,84]	91 [85,96]
segmentation	56 [41,73]	56 [43,68]	89 [82,95]	78 [67,88]	74 [65,83]	90 [85,95]
nnUNet	76 [60,89]	69 [56,80]	88 [80,95]	70 [57,82]	69 [59,79]	86 [79,92]
R-Super†	77 [63,90]	70 [58,81]	93 [87,97]	82 [70,91]	76 [67,85]	95 [92,98]

AI	gallbladder malignant tumor			uterus malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	n/a	n/a	n/a	n/a	n/a	n/a
R-Super	n/a	n/a	n/a	n/a	n/a	n/a
many reports, few masks training						
CLIP	n/a	n/a	n/a	n/a	n/a	n/a
MTL	n/a	n/a	n/a	n/a	n/a	n/a
ModelsGenesis	n/a	n/a	n/a	n/a	n/a	n/a
segmentation	n/a	n/a	n/a	n/a	n/a	n/a
nnUNet	n/a	n/a	n/a	n/a	n/a	n/a
R-Super†	n/a	n/a	n/a	n/a	n/a	n/a

AI	prostate malignant tumor			adrenal malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	n/a	n/a	n/a	29 [18,42]	37 [24,49]	64 [56,73]
R-Super	n/a	n/a	n/a	95 [89,100]	85 [78,91]	97 [95,99]
many reports, few masks training						
CLIP	n/a	n/a	n/a	95 [89,100]	85 [78,91]	97 [96,99]
MTL	n/a	n/a	n/a	97 [91,100]	86 [79,92]	98 [96,99]
ModelsGenesis	n/a	n/a	n/a	96 [91,100]	86 [79,92]	97 [94,99]
segmentation	n/a	n/a	n/a	97 [91,100]	86 [79,92]	98 [96,99]
nnUNet	n/a	n/a	n/a	90 [81,97]	82 [74,88]	92 [88,97]
R-Super†	n/a	n/a	n/a	93 [86,98]	84 [77,90]	96 [92,99]

AI	spleen malignant tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	14 [8,21]	22 [12,31]	54 [47,61]	22 [17,27]	29 [22,35]	66 [62,70]
R-Super	72 [63,80]	77 [69,83]	87 [82,92]	85 [81,90]	79 [75,83]	93 [90,95]
many reports, few masks training						
CLIP	41 [32,51]	53 [43,62]	80 [74,85]	66 [60,71]	66 [60,71]	87 [83,90]
MTL	66 [57,75]	73 [65,80]	88 [83,92]	81 [76,86]	77 [72,81]	92 [89,94]
ModelsGenesis	32 [23,40]	43 [33,52]	81 [76,86]	67 [62,73]	66 [60,71]	89 [86,91]
segmentation	41 [32,52]	52 [43,62]	85 [80,90]	68 [62,74]	67 [61,73]	90 [88,93]
nnUNet	55 [46,65]	65 [56,73]	79 [73,85]	72 [67,78]	71 [66,76]	86 [83,89]
R-Super†	74 [65,82]	78 [71,84]	90 [86,94]	81 [76,86]	77 [72,82]	94 [91,95]

Table D35: **Malignant tumor detection, external validation (Basel, Switzerland): comparison to alternative methods trained on the same dataset.** This table evaluates tumor detection stratified for malignant tumors (malignant vs no-tumor). All models were trained on the R-Super train set (USA), and evaluated on the Basel test set (external validation, Switzerland). See dataset details in Tab. 1, A1. All models were evaluated at 90% specificity. We compare multiple training methods: R-Super†, MTL, and CLIP were trained on all available CT-Report plus CT-Mask pairs; segmentation and nnU-Net on CT-Mask pairs only; R-Super and classification on CT-Report only; ModelsGenesis on all CT scans and all CT-Mask pairs. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model (separately for models trained without masks). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% CI (1,000 bootstrap samples).

AI	bladder malignant tumor			esophagus malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	30 [20,40]	43 [32,55]	82 [73,90]	68 [48,85]	73 [57,85]	84 [73,93]
R-Super	81 [72,90]	86 [80,92]	93 [88,97]	64 [44,82]	71 [54,83]	82 [69,92]
many reports, few masks training						
CLIP	86 [77,93]	89 [83,94]	92 [87,96]	52 [32,70]	61 [42,76]	77 [63,87]
MTL	83 [75,92]	88 [82,93]	93 [89,97]	64 [45,81]	71 [54,83]	78 [65,90]
ModelsGenesis	74 [64,84]	82 [75,89]	93 [88,97]	57 [38,74]	65 [47,79]	73 [58,87]
segmentation	77 [68,86]	85 [78,90]	89 [83,94]	62 [41,82]	68 [50,82]	80 [67,91]
nnUNet	80 [69,89]	86 [78,92]	90 [84,95]	64 [46,82]	71 [55,84]	82 [71,92]
R-Super†	88 [81,95]	91 [85,95]	94 [89,97]	64 [45,81]	71 [55,83]	89 [79,97]

AI	gallbladder malignant tumor			uterus malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	0 [0,0]	0 [0,0]	28 [18,40]	93 [83,100]	89 [79,96]	97 [93,99]
R-Super	24 [5,47]	31 [8,53]	69 [54,84]	93 [83,100]	89 [79,96]	95 [87,100]
many reports, few masks training						
CLIP	69 [45,92]	69 [48,85]	92 [85,98]	100 [100,100]	92 [84,98]	100 [99,100]
MTL	59 [35,83]	67 [44,84]	90 [82,97]	100 [100,100]	95 [89,100]	100 [99,100]
ModelsGenesis	53 [28,79]	58 [33,76]	83 [71,93]	100 [100,100]	92 [84,98]	99 [97,100]
segmentation	56 [31,80]	60 [36,78]	81 [67,93]	100 [100,100]	94 [86,100]	100 [99,100]
nnUNet	59 [33,82]	62 [38,80]	84 [71,95]	97 [89,100]	91 [82,97]	98 [94,100]
R-Super†	76 [56,96]	74 [56,90]	95 [89,99]	100 [100,100]	92 [84,98]	100 [100,100]

AI	prostate malignant tumor			adrenal malignant tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	43 [25,62]	53 [34,70]	82 [72,91]	5 [3,8]	10 [6,14]	48 [40,56]
R-Super	71 [54,88]	76 [60,87]	87 [77,95]	32 [28,37]	49 [43,54]	83 [78,88]
many reports, few masks training						
CLIP	41 [23,58]	51 [31,68]	81 [71,90]	77 [73,81]	87 [84,89]	90 [87,93]
MTL	61 [41,80]	69 [52,83]	84 [74,93]	76 [72,80]	86 [83,88]	87 [83,90]
ModelsGenesis	54 [36,72]	62 [44,77]	80 [68,89]	76 [72,80]	86 [83,88]	90 [86,93]
segmentation	40 [21,61]	50 [29,67]	79 [68,90]	81 [77,85]	89 [86,91]	89 [86,93]
nnUNet	56 [35,74]	64 [44,78]	74 [62,85]	73 [69,77]	84 [81,87]	86 [82,89]
R-Super†	64 [45,83]	71 [54,83]	86 [74,94]	78 [73,81]	87 [84,89]	86 [82,90]

AI	spleen malignant tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	4 [1,9]	8 [2,16]	51 [41,62]	35 [30,39]	40 [35,43]	67 [64,72]
R-Super	52 [41,62]	66 [55,75]	81 [74,88]	60 [55,65]	67 [61,71]	84 [81,87]
many reports, few masks training						
CLIP	50 [39,60]	64 [54,73]	76 [68,83]	68 [62,73]	73 [68,78]	87 [84,89]
MTL	26 [17,36]	40 [28,50]	70 [61,79]	67 [62,72]	74 [68,78]	86 [83,89]
ModelsGenesis	34 [24,44]	49 [38,58]	72 [64,80]	64 [58,70]	71 [65,75]	84 [81,87]
segmentation	34 [25,45]	48 [38,60]	74 [65,83]	64 [59,70]	71 [65,75]	85 [81,88]
nnUNet	53 [42,64]	67 [57,76]	74 [66,82]	69 [63,75]	75 [70,80]	84 [81,87]
R-Super†	67 [57,78]	78 [70,85]	86 [79,92]	77 [72,82]	80 [76,84]	91 [88,93]

D.7.3 Small Tumor Detection

Table D36: **Small tumor detection, internal validation (UCSF, USA): comparison to alternative methods trained on the same dataset in detecting small tumors.** This table evaluates tumor detection (malignant + benign vs no-tumor), considering only small tumors (diameter $\leq$ 2 cm). All models were trained on the R-Super train set (UCSF + Stanford), and evaluated on the UCSF test set (internal validation). See dataset details in Tab. 1, A1. All models were evaluated at 90% specificity. We compare multiple training methods: R-Super†, MTL, and CLIP were trained on all available CT-Report plus CT-Mask pairs; segmentation and nnU-Net on CT-Mask pairs only; R-Super and classification on CT-Report only; ModelsGenesis on all CT scans and all CT-Mask pairs. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model (separately for models trained without masks). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% CI (1,000 bootstrap samples).

AI	bladder small tumor			esophagus small tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	15 [9,22]	24 [14,33]	61 [53,68]	64 [52,76]	71 [60,79]	88 [83,93]
R-Super	33 [24,41]	46 [36,55]	67 [60,74]	86 [76,94]	85 [77,91]	93 [89,96]
many reports, few masks training						
CLIP	14 [8,21]	23 [14,32]	66 [59,74]	81 [71,90]	82 [74,88]	88 [81,94]
MTL	20 [13,28]	30 [20,40]	71 [64,78]	81 [71,90]	82 [74,89]	91 [85,96]
ModelsGenesis	24 [16,32]	36 [25,46]	70 [63,76]	73 [62,83]	77 [67,84]	86 [80,92]
segmentation	23 [16,32]	35 [25,45]	70 [63,76]	81 [70,90]	82 [73,88]	88 [81,94]
nnUNet	22 [14,30]	33 [23,43]	66 [58,72]	84 [74,92]	84 [76,90]	91 [86,95]
R-Super†	50 [40,59]	62 [53,70]	74 [67,81]	89 [81,97]	86 [79,92]	91 [84,97]

AI	gallbladder small tumor			uterus small tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	8 [2,14]	12 [3,22]	53 [44,62]	49 [36,61]	59 [46,69]	78 [69,86]
R-Super	31 [20,42]	42 [28,53]	70 [62,77]	68 [56,79]	73 [63,82]	86 [79,92]
many reports, few masks training						
CLIP	31 [19,41]	42 [28,52]	66 [58,74]	41 [29,54]	51 [39,63]	78 [71,85]
MTL	34 [22,46]	45 [31,57]	68 [60,76]	49 [37,62]	59 [46,70]	85 [79,91]
ModelsGenesis	31 [19,42]	41 [28,54]	68 [60,75]	42 [29,55]	52 [38,65]	79 [72,86]
segmentation	42 [29,53]	52 [40,63]	71 [63,78]	41 [29,53]	51 [38,63]	81 [75,88]
nnUNet	28 [17,39]	39 [25,51]	60 [51,68]	45 [33,58]	56 [43,67]	80 [73,86]
R-Super†	51 [38,63]	61 [49,71]	76 [68,84]	63 [52,75]	69 [59,79]	81 [73,88]

AI	prostate small tumor			adrenal small tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	10 [0,22]	14 [0,28]	65 [53,76]	5 [2,8]	9 [5,14]	42 [36,48]
R-Super	45 [27,62]	50 [32,65]	72 [59,82]	85 [81,89]	90 [87,92]	93 [90,96]
many reports, few masks training						
CLIP	19 [7,34]	26 [9,42]	60 [49,70]	63 [57,69]	76 [71,80]	87 [84,91]
MTL	37 [19,55]	42 [23,59]	69 [58,79]	73 [68,78]	82 [79,86]	89 [85,92]
ModelsGenesis	19 [5,35]	24 [6,41]	63 [52,74]	65 [59,71]	77 [72,81]	87 [84,90]
segmentation	29 [14,46]	35 [17,51]	69 [59,79]	76 [71,81]	84 [81,88]	89 [86,92]
nnUNet	23 [9,39]	30 [12,46]	67 [56,77]	64 [58,69]	76 [72,80]	81 [77,85]
R-Super†	48 [30,67]	53 [35,68]	81 [71,89]	82 [78,87]	88 [85,91]	90 [86,93]

AI	spleen small tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	11 [6,16]	19 [11,26]	47 [40,54]	23 [20,26]	30 [26,33]	62 [58,65]
R-Super	78 [70,84]	84 [79,88]	86 [80,90]	61 [57,65]	67 [63,71]	81 [78,84]
many reports, few masks training						
CLIP	23 [16,30]	36 [27,44]	71 [65,77]	39 [35,43]	48 [44,52]	74 [71,77]
MTL	49 [42,58]	63 [56,70]	78 [72,83]	49 [45,53]	58 [53,62]	79 [76,81]
ModelsGenesis	24 [17,31]	37 [28,46]	74 [67,79]	40 [36,44]	49 [44,53]	75 [72,78]
segmentation	39 [31,46]	53 [45,61]	79 [72,84]	47 [43,51]	56 [52,60]	78 [75,81]
nnUNet	33 [25,41]	47 [38,55]	70 [64,77]	43 [39,47]	52 [48,56]	73 [71,76]
R-Super†	77 [70,84]	84 [78,88]	86 [81,91]	66 [62,70]	72 [68,75]	83 [80,85]

Table D37: **Small tumor detection, external validation (Stanford, USA): comparison to alternative methods trained on the same dataset in detecting small tumors.** This table evaluates tumor detection (malignant + benign vs no-tumor), considering only small tumors (diameter $\leq$ 2 cm). All models were trained on the UCSF train set (USA), and evaluated on the Stanford test set (external validation, USA). See dataset details in Tab. 1, A1. All models were evaluated at 90% specificity. We compare multiple training methods: **R-Super**†, MTL, and CLIP were trained on all available CT-Report plus CT-Mask pairs; segmentation and nnU-Net on CT-Mask pairs only; **R-Super** and classification on CT-Report only; ModelsGenesis on all CT scans and all CT-Mask pairs. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model (separately for models trained without masks). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% CI (1,000 bootstrap samples). n/a: we skipped organs with less than 10 positive scans in the dataset.

AI	bladder small tumor			esophagus small tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	2 [0,8]	4 [0,12]	45 [35,55]	n/a	n/a	n/a
R-Super	51 [36,66]	57 [42,70]	81 [72,88]	n/a	n/a	n/a
many reports, few masks training						
CLIP	29 [16,44]	38 [23,52]	74 [65,82]	n/a	n/a	n/a
MTL	76 [62,88]	76 [65,86]	84 [77,92]	n/a	n/a	n/a
ModelsGenesis	32 [18,46]	40 [24,54]	70 [61,79]	n/a	n/a	n/a
segmentation	36 [21,52]	44 [28,59]	72 [63,80]	n/a	n/a	n/a
nnUNet	15 [5,28]	21 [7,35]	59 [48,70]	n/a	n/a	n/a
R-Super†	73 [59,87]	74 [62,83]	84 [74,92]	n/a	n/a	n/a

AI	gallbladder small tumor			uterus small tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	12 [3,24]	18 [4,32]	46 [36,56]	6 [0,14]	10 [0,20]	70 [61,78]
R-Super	49 [33,64]	56 [41,69]	76 [66,85]	60 [45,74]	65 [52,76]	87 [80,93]
many reports, few masks training						
CLIP	46 [31,61]	54 [38,68]	71 [62,80]	23 [12,36]	32 [17,45]	75 [68,83]
MTL	27 [13,41]	35 [18,49]	61 [52,69]	30 [17,44]	39 [24,53]	79 [72,86]
ModelsGenesis	42 [26,58]	49 [34,63]	71 [62,79]	28 [15,41]	37 [21,50]	74 [66,81]
segmentation	51 [36,66]	60 [46,72]	73 [65,81]	28 [16,40]	37 [22,50]	78 [70,85]
nnUNet	0 [0,0]	0 [0,0]	50 [50,50]	48 [33,62]	56 [42,67]	79 [70,86]
R-Super†	68 [54,82]	70 [57,81]	88 [82,94]	60 [45,73]	65 [52,76]	84 [78,90]

AI	prostate small tumor			adrenal small tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	n/a	n/a	n/a	16 [12,19]	26 [21,32]	56 [50,62]
R-Super	n/a	n/a	n/a	87 [84,90]	92 [89,94]	94 [92,96]
many reports, few masks training						
CLIP	n/a	n/a	n/a	68 [64,73]	80 [76,83]	82 [78,84]
MTL	n/a	n/a	n/a	77 [73,82]	86 [83,89]	86 [83,89]
ModelsGenesis	n/a	n/a	n/a	66 [60,70]	78 [75,82]	81 [78,84]
segmentation	n/a	n/a	n/a	70 [65,74]	81 [78,84]	84 [81,86]
nnUNet	n/a	n/a	n/a	50 [44,55]	65 [60,70]	73 [69,77]
R-Super†	n/a	n/a	n/a	83 [79,86]	89 [87,91]	92 [89,94]

AI	spleen small tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	10 [6,14]	17 [10,24]	51 [44,57]	9 [7,13]	15 [11,20]	54 [49,58]
R-Super	70 [63,77]	80 [74,84]	83 [79,88]	63 [58,69]	70 [65,74]	84 [81,87]
many reports, few masks training						
CLIP	55 [47,62]	68 [61,74]	75 [70,79]	44 [40,50]	54 [49,59]	75 [72,79]
MTL	64 [57,72]	76 [70,82]	80 [76,84]	55 [50,59]	63 [58,67]	78 [76,81]
ModelsGenesis	63 [56,70]	75 [69,80]	79 [74,83]	46 [41,52]	56 [50,61]	75 [72,78]
segmentation	61 [54,68]	73 [68,78]	78 [73,82]	49 [44,54]	59 [54,64]	77 [74,80]
nnUNet	50 [43,57]	64 [57,70]	76 [71,81]	32 [28,37]	41 [37,45]	68 [64,70]
R-Super†	71 [64,78]	80 [75,85]	84 [79,88]	71 [66,76]	76 [71,80]	86 [84,89]

Table D38: **Small tumor detection, external validation (Medipol, Turkey): comparison to alternative methods trained on the same dataset in detecting small tumors.** This table evaluates tumor detection (malignant + benign vs no-tumor), considering only small tumors (diameter ≤ 2 cm). All models were trained on the R-Super train set (USA), and evaluated on the Medipol test set (external validation, Turkey). See dataset details in Tab. 1, A1. All models were evaluated at 90% specificity. We compare multiple training methods: **R-Super**†, MTL, and CLIP were trained on all available CT-Report plus CT-Mask pairs; segmentation and nnU-Net on CT-Mask pairs only; **R-Super** and classification on CT-Report only; ModelsGenesis on all CT scans and all CT-Mask pairs. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model (separately for models trained without masks). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% CI (1,000 bootstrap samples). n/a: we skipped organs with less than 10 positive scans in the dataset.

AI	bladder small tumor			esophagus small tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	4 [0,10]	6 [0,14]	57 [50,64]	30 [16,43]	36 [21,49]	72 [62,80]
R-Super	37 [24,51]	44 [30,56]	80 [74,85]	77 [64,89]	72 [60,81]	93 [89,97]
many reports, few masks training						
CLIP	2 [0,6]	3 [0,9]	63 [55,70]	57 [42,73]	58 [44,70]	83 [75,90]
MTL	43 [30,56]	49 [36,60]	70 [62,78]	64 [49,77]	63 [50,73]	86 [77,92]
ModelsGenesis	11 [3,21]	16 [5,28]	62 [54,71]	54 [39,69]	56 [42,68]	86 [78,92]
segmentation	13 [5,22]	18 [7,29]	68 [61,75]	68 [54,82]	66 [55,77]	88 [82,93]
nnUNet	43 [30,56]	50 [37,61]	69 [60,77]	62 [47,77]	61 [47,72]	78 [68,87]
R-Super†	37 [24,50]	44 [31,56]	80 [73,85]	64 [49,78]	63 [50,74]	90 [84,95]

AI	gallbladder small tumor			uterus small tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	4 [0,9]	6 [0,14]	32 [25,39]	17 [0,43]	13 [0,29]	72 [60,84]
R-Super	39 [28,49]	48 [37,58]	68 [61,76]	8 [0,29]	7 [0,21]	83 [76,88]
many reports, few masks training						
CLIP	37 [26,49]	47 [34,57]	79 [73,85]	8 [0,29]	7 [0,22]	74 [62,85]
MTL	32 [22,43]	41 [30,53]	75 [69,81]	8 [0,29]	7 [0,21]	82 [75,89]
ModelsGenesis	19 [10,30]	27 [15,40]	74 [67,80]	8 [0,27]	7 [0,19]	66 [56,77]
segmentation	41 [30,53]	50 [39,60]	79 [73,84]	8 [0,30]	7 [0,21]	75 [66,83]
nnUNet	18 [9,27]	26 [14,37]	57 [50,65]	17 [0,38]	13 [0,30]	58 [43,76]
R-Super†	71 [60,81]	74 [65,81]	80 [73,87]	17 [0,43]	15 [0,35]	80 [72,88]

AI	prostate small tumor			adrenal small tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	n/a	n/a	n/a	18 [11,25]	27 [17,36]	64 [57,70]
R-Super	n/a	n/a	n/a	94 [89,98]	90 [85,94]	96 [93,98]
many reports, few masks training						
CLIP	n/a	n/a	n/a	71 [63,79]	77 [70,82]	91 [88,94]
MTL	n/a	n/a	n/a	82 [75,89]	84 [78,89]	93 [89,96]
ModelsGenesis	n/a	n/a	n/a	80 [72,87]	82 [76,87]	92 [88,95]
segmentation	n/a	n/a	n/a	86 [79,92]	86 [80,90]	94 [91,96]
nnUNet	n/a	n/a	n/a	70 [61,78]	75 [69,81]	82 [77,87]
R-Super†	n/a	n/a	n/a	87 [80,92]	86 [81,90]	91 [87,95]

AI	spleen small tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	12 [7,17]	19 [12,27]	54 [47,60]	14 [10,19]	18 [13,22]	58 [55,62]
R-Super	70 [63,78]	77 [71,83]	85 [81,89]	54 [50,59]	56 [52,60]	84 [82,86]
many reports, few masks training						
CLIP	6 [3,10]	10 [5,17]	66 [60,72]	30 [26,35]	33 [29,38]	76 [73,79]
MTL	28 [21,35]	40 [31,48]	76 [71,81]	43 [38,48]	47 [42,52]	80 [78,83]
ModelsGenesis	4 [1,7]	6 [1,12]	74 [69,80]	29 [25,34]	32 [28,36]	76 [72,79]
segmentation	2 [0,5]	4 [0,9]	75 [69,81]	36 [32,42]	38 [34,42]	80 [77,82]
nnUNet	26 [19,33]	38 [29,46]	71 [66,78]	39 [34,44]	44 [39,48]	69 [66,74]
R-Super†	57 [49,65]	67 [60,74]	81 [76,86]	55 [50,61]	58 [53,63]	84 [81,86]

Table D39: **Small tumor detection, external validation (Basel, Switzerland): comparison to alternative methods trained on the same dataset in detecting small tumors.** This table evaluates tumor detection (malignant + benign vs no-tumor), considering only small tumors (diameter $\leq$ 2 cm). All models were trained on the R-Super train set (USA), and evaluated on the Basel test set (external validation, Switzerland). See dataset details in Tab. 1, A1. All models were evaluated at 90% specificity. We compare multiple training methods: **R-Super**†, MTL, and CLIP were trained on all available CT-Report plus CT-Mask pairs; segmentation and nnU-Net on CT-Mask pairs only; **R-Super** and classification on CT-Report only; ModelsGenesis on all CT scans and all CT-Mask pairs. Blue indicates statistical significance ($p < 0.05$), comparing the best-performing model with the second-best model (separately for models trained without masks). We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% CI (1,000 bootstrap samples). n/a: we skipped organs with less than 10 positive scans in the dataset.

AI	bladder small tumor			esophagus small tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	19 [7,33]	29 [12,46]	80 [70,88]	n/a	n/a	n/a
R-Super	54 [39,70]	64 [50,77]	83 [74,91]	n/a	n/a	n/a
many reports, few masks training						
CLIP	78 [64,91]	81 [70,90]	90 [82,96]	n/a	n/a	n/a
MTL	81 [68,93]	83 [72,91]	88 [81,95]	n/a	n/a	n/a
ModelsGenesis	57 [39,73]	67 [51,79]	88 [81,95]	n/a	n/a	n/a
segmentation	68 [52,82]	76 [64,86]	85 [77,94]	n/a	n/a	n/a
nnUNet	76 [62,89]	80 [69,89]	84 [76,92]	n/a	n/a	n/a
R-Super†	89 [79,98]	88 [80,94]	92 [84,98]	n/a	n/a	n/a

AI	gallbladder small tumor			uterus small tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	n/a	n/a	n/a	67 [38,92]	64 [38,84]	85 [70,97]
R-Super	n/a	n/a	n/a	50 [20,82]	52 [24,75]	78 [54,97]
many reports, few masks training						
CLIP	n/a	n/a	n/a	75 [46,100]	69 [42,87]	86 [69,100]
MTL	n/a	n/a	n/a	83 [60,100]	80 [59,95]	91 [78,100]
ModelsGenesis	n/a	n/a	n/a	75 [50,100]	69 [44,88]	85 [70,99]
segmentation	n/a	n/a	n/a	90 [67,100]	82 [59,96]	95 [82,100]
nnUNet	n/a	n/a	n/a	67 [36,91]	64 [36,83]	81 [65,95]
R-Super†	n/a	n/a	n/a	83 [58,100]	74 [52,91]	90 [75,100]

AI	prostate small tumor			adrenal small tumor		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	62 [33,90]	62 [36,80]	84 [69,95]	14 [9,20]	24 [15,33]	53 [45,63]
R-Super	62 [33,88]	62 [35,80]	82 [66,95]	28 [21,36]	43 [34,52]	79 [73,85]
many reports, few masks training						
CLIP	67 [38,93]	64 [38,83]	85 [69,97]	69 [61,77]	80 [75,86]	85 [79,90]
MTL	69 [43,93]	69 [46,88]	86 [73,97]	68 [61,76]	80 [74,85]	84 [78,90]
ModelsGenesis	54 [25,82]	56 [28,77]	84 [71,96]	71 [63,78]	81 [76,86]	86 [80,92]
segmentation	75 [46,100]	69 [44,88]	90 [78,99]	80 [72,86]	87 [82,91]	87 [81,93]
nnUNet	43 [15,71]	48 [19,71]	69 [54,83]	59 [51,68]	74 [66,80]	77 [70,84]
R-Super†	62 [33,88]	62 [33,80]	84 [68,96]	71 [63,78]	81 [75,86]	84 [78,90]

AI	spleen small tumor			**Average**		
	sen.	F1	AUC	sen.	F1	AUC
report only training						
classification	4 [1,8]	7 [2,14]	45 [36,57]	33 [24,41]	37 [28,44]	69 [64,75]
R-Super	52 [43,62]	67 [58,74]	75 [67,82]	49 [40,58]	58 [49,65]	79 [73,85]
many reports, few masks training						
CLIP	9 [4,15]	16 [8,25]	55 [44,64]	60 [50,67]	62 [54,69]	80 [75,85]
MTL	3 [0,6]	5 [0,11]	46 [37,56]	61 [53,68]	64 [56,69]	79 [74,83]
ModelsGenesis	10 [5,16]	18 [9,27]	51 [41,60]	53 [45,62]	58 [50,65]	79 [74,83]
segmentation	4 [1,8]	7 [2,14]	52 [41,62]	63 [56,70]	64 [57,70]	82 [77,86]
nnUNet	11 [6,16]	18 [10,27]	52 [42,61]	51 [42,59]	57 [49,64]	73 [68,78]
R-Super†	38 [28,47]	53 [43,63]	76 [68,84]	68 [60,76]	72 [64,77]	85 [80,89]

D.8 Comparison to Radiologists

Here, we compare R-Super to radiologists at the probable confidence level. As described in Section 2.4, for each CT scan, radiologists indicated which organs probably or certainly contained a tumor, and which probably or certainly contained a malignant tumor. Radiologists were instructed to mark at the probable level any tumor they would mention in a radiology report, simulating which tumors would be detected in clinical practice. All tumors marked at the certain level were also counted as positives for the probable level. In the tables below, R-Super is evaluated at specificity matched to the radiologists'.

Table D40: **Experience of radiologists in the reader study**. The number of CT scans read per year are approximate, estimated by the radiologists. Experience time includes residency: the 3 radiologists with 2 years of experience are residents. The other 3 are board-certified radiologists. We grouped radiologists in pairs. Each radiologist in a pair read half of the CT scans in the dataset. These two dataset halves were randomly split, keeping tumor proportions the same.

radiologist ID	group	experience (years)	CTs read per year	training / expertise
1	group A	2	700	chest and abdominal radiology
2	group A	2	1,100	chest and abdominal radiology
3	group B	2	800	chest and abdominal radiology
4	group B	4	3,000	abdominal radiology
5	group C	8	2,000	chest and abdominal radiology
6	group C	5	2,000	abdominal radiology

Table D41: **Reader Study: tumor detection.** R-Super surpassed the radiologist average in sensitivity for every tumor type except uterus, achieving +22% average sensitivity at approximately matched specificity. Here, we show results for tumor detection (benign + malignant vs no-tumor). Blue indicates statistical significance ($p < 0.05$), comparing the AI model with radiologist average. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	36 [26,46]	98 [95,100]	52 [40,61]	67 [62,72]	55 [44,66]	97 [94,99]	69 [58,77]	76 [71,82]
rad. group B	37 [28,46]	98 [95,100]	53 [42,62]	68 [63,72]	55 [44,65]	97 [94,99]	69 [59,77]	76 [71,82]
rad. group C	23 [15,32]	100 [100,100]	37 [26,48]	61 [57,66]	4 [0,8]	100 [100,100]	7 [0,16]	52 [50,54]
radiologist average	32 [27,37]	98 [97,100]	48 [41,53]	65 [63,68]	38 [32,44]	98 [97,99]	54 [47,60]	68 [65,71]
R-Super†	48 [38,58]	98 [95,100]	64 [54,72]	88 [82,92]	67 [57,77]	98 [95,100]	79 [70,86]	92 [87,96]

	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	47 [37,58]	95 [90,98]	60 [50,70]	72 [65,77]	74 [66,82]	85 [79,91]	77 [70,83]	82 [77,87]
rad. group B	49 [38,59]	95 [91,98]	62 [52,71]	72 [66,78]	75 [66,83]	84 [78,90]	77 [70,83]	82 [77,87]
rad. group C	24 [15,34]	100 [100,100]	39 [26,50]	62 [57,67]	69 [61,78]	98 [96,100]	81 [74,87]	84 [80,88]
radiologist average	40 [34,46]	96 [95,98]	55 [48,61]	69 [66,72]	73 [68,78]	89 [86,92]	78 [74,82]	83 [80,86]
R-Super†	69 [59,78]	96 [93,99]	79 [71,85]	90 [85,94]	63 [54,72]	89 [83,94]	71 [64,78]	82 [77,87]

	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	38 [27,48]	96 [92,99]	52 [40,62]	67 [61,72]	8 [4,13]	100 [100,100]	14 [7,23]	54 [52,56]
rad. group B	38 [27,49]	96 [92,99]	52 [40,63]	67 [62,73]	5 [2,8]	100 [100,100]	9 [3,16]	52 [51,54]
rad. group C	14 [7,23]	100 [100,100]	25 [14,37]	57 [54,61]	48 [40,56]	98 [95,100]	64 [56,71]	73 [68,77]
radiologist average	30 [24,36]	97 [95,98]	44 [37,51]	64 [61,66]	20 [17,24]	99 [98,100]	34 [28,39]	60 [58,62]
R-Super†	54 [43,66]	97 [94,99]	68 [58,77]	90 [85,95]	54 [46,63]	99 [98,100]	70 [62,77]	92 [89,95]

	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	43 [33,52]	96 [93,99]	58 [49,67]	70 [65,75]	43 [39,46]	95 [94,96]	55 [51,58]	70 [68,72]
rad. group B	43 [33,52]	97 [94,99]	59 [49,67]	70 [65,75]	43 [39,46]	95 [94,96]	54 [51,57]	70 [68,72]
rad. group C	35 [26,44]	100 [100,100]	52 [41,61]	67 [63,72]	31 [28,34]	99 [99,100]	44 [40,47]	65 [64,67]
radiologist average	40 [35,46]	98 [96,99]	56 [51,62]	69 [66,72]	39 [37,41]	96 [96,97]	53 [50,55]	68 [67,69]
R-Super†	71 [62,79]	98 [95,100]	82 [75,87]	86 [80,91]	61 [57,65]	96 [95,98]	73 [70,76]	88 [87,90]

Table D42: **Reader Study: malignant tumor detection.** R-Super surpassed the radiologist average in sensitivity for all tumor types, with gains ranging up to +48%, and +27% on average. Here, we show results for malignant tumor detection (malignant vs no-tumor). Blue indicates statistical significance ($p < 0.05$), comparing the AI model with radiologist average. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder malignant tumor				esophagus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	57 [40,75]	98 [95,100]	69 [53,82]	78 [69,87]	67 [53,81]	97 [94,99]	76 [64,86]	83 [76,90]
rad. group B	57 [40,74]	98 [95,100]	69 [53,82]	78 [69,86]	67 [52,81]	97 [94,99]	76 [64,86]	83 [75,90]
rad. group C	37 [22,55]	100 [100,100]	54 [36,71]	69 [61,77]	2 [0,8]	100 [100,100]	4 [0,15]	51 [50,54]
radiologist average	50 [41,60]	98 [97,100]	65 [55,72]	75 [70,80]	46 [37,55]	98 [97,99]	60 [52,68]	72 [68,77]
R-Super†	71 [56,85]	98 [95,100]	79 [67,89]	95 [89,99]	72 [58,85]	98 [95,100]	80 [69,89]	96 [93,98]

	gallbladder malignant tumor				uterus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	68 [48,87]	95 [90,98]	69 [51,82]	82 [72,92]	76 [61,89]	85 [79,91]	64 [50,75]	84 [75,91]
rad. group B	68 [50,85]	95 [91,98]	69 [53,82]	82 [72,92]	76 [61,89]	84 [78,90]	63 [50,75]	83 [75,91]
rad. group C	32 [13,52]	100 [100,100]	48 [23,68]	66 [56,76]	79 [64,92]	98 [96,100]	85 [74,94]	89 [82,96]
radiologist average	56 [44,68]	96 [95,98]	64 [55,74]	77 [71,83]	77 [68,84]	89 [86,92]	70 [62,76]	85 [80,89]
R-Super†	96 [88,100]	96 [93,99]	89 [79,96]	99 [98,100]	79 [64,92]	89 [83,94]	70 [58,81]	86 [78,93]

	prostate malignant tumor				adrenal malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	42 [23,61]	96 [92,99]	51 [31,68]	69 [60,79]	13 [3,24]	100 [100,100]	23 [6,39]	56 [52,62]
rad. group B	42 [26,63]	96 [92,99]	51 [33,68]	69 [61,80]	8 [0,16]	100 [100,100]	14 [0,28]	54 [50,58]
rad. group C	27 [10,45]	100 [100,100]	42 [18,62]	64 [55,73]	36 [21,51]	98 [95,100]	50 [32,65]	67 [59,75]
radiologist average	37 [27,48]	97 [95,98]	49 [38,59]	67 [62,73]	19 [12,26]	99 [98,100]	31 [21,41]	59 [56,62]
R-Super†	65 [46,83]	97 [94,99]	72 [56,85]	92 [83,98]	67 [50,81]	99 [98,100]	79 [66,88]	94 [89,99]

	spleen malignant tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	54 [32,77]	96 [93,99]	62 [41,78]	76 [64,88]	54 [48,60]	95 [94,96]	59 [53,64]	76 [72,79]
rad. group B	54 [32,76]	97 [94,99]	63 [42,79]	76 [65,87]	53 [47,59]	95 [94,96]	58 [52,63]	75 [72,78]
rad. group C	41 [21,64]	100 [100,100]	58 [34,78]	70 [60,82]	36 [30,43]	99 [99,100]	49 [41,55]	68 [65,71]
radiologist average	50 [38,61]	98 [96,99]	61 [50,71]	74 [68,80]	48 [44,52]	96 [96,97]	57 [53,61]	73 [71,75]
R-Super†	77 [58,94]	98 [95,100]	81 [67,92]	94 [88,99]	75 [69,81]	96 [95,98]	79 [74,83]	94 [92,96]

Table D43: **Reader Study: malignancy identification.** Here, we show results for malignancy identification, i.e., distinguishing malignant tumors from benign tumors plus no-tumor cases. R-Super surpassed the radiologist average in sensitivity for 5 tumor types at approximately matched specificity, achieving +14% sensitivity on average. Blue indicates statistical significance ($p < 0.05$), comparing the AI model with radiologist average. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder malignancy id.				esophagus malignancy id.			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	46 [30,64]	97 [95,99]	58 [41,73]	72 [64,81]	63 [48,77]	93 [88,96]	67 [54,77]	79 [71,86]
rad. group B	43 [27,61]	97 [95,99]	56 [38,71]	70 [62,79]	65 [50,79]	93 [88,97]	68 [56,80]	80 [72,87]
rad. group C	37 [22,55]	99 [98,100]	53 [35,69]	68 [60,77]	2 [0,8]	99 [98,100]	4 [0,14]	51 [49,54]
radiologist average	42 [32,52]	98 [97,99]	56 [46,65]	70 [65,75]	43 [35,51]	95 [93,97]	54 [46,62]	70 [65,74]
R-Super†	57 [40,73]	98 [96,100]	69 [54,81]	93 [86,98]	63 [48,77]	95 [92,99]	70 [57,81]	90 [85,95]

	gallbladder malignancy id.				uterus malignancy id.			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	56 [36,76]	93 [88,96]	56 [38,71]	76 [65,87]	64 [46,79]	76 [69,81]	42 [29,54]	72 [63,82]
rad. group B	56 [35,76]	94 [90,97]	58 [39,74]	76 [65,86]	64 [47,79]	77 [71,82]	42 [30,53]	73 [63,82]
rad. group C	28 [11,48]	100 [100,100]	44 [19,64]	64 [55,74]	73 [57,87]	97 [94,99]	76 [63,87]	85 [77,93]
radiologist average	47 [36,57]	96 [94,97]	54 [43,63]	72 [66,78]	67 [57,76]	83 [80,86]	50 [43,57]	77 [72,82]
R-Super†	72 [52,89]	96 [92,99]	74 [57,86]	95 [90,99]	64 [47,81]	83 [78,88]	49 [36,60]	77 [66,86]

	prostate malignancy id.				adrenal malignancy id.			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	38 [20,58]	94 [90,98]	46 [25,63]	67 [57,77]	13 [3,24]	98 [96,100]	21 [6,36]	55 [50,61]
rad. group B	38 [21,59]	94 [90,98]	46 [26,63]	67 [58,78]	8 [0,16]	98 [96,100]	13 [0,27]	53 [49,57]
rad. group C	23 [8,39]	100 [100,100]	38 [14,56]	62 [54,70]	28 [14,43]	97 [94,99]	40 [23,56]	63 [56,71]
radiologist average	33 [23,43]	96 [94,98]	43 [31,54]	65 [60,70]	16 [10,23]	98 [96,99]	26 [17,35]	57 [54,61]
R-Super†	65 [46,83]	96 [92,99]	69 [52,82]	91 [83,97]	15 [5,28]	98 [96,100]	24 [10,40]	79 [69,88]

	spleen malignancy id.				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	32 [13,55]	95 [92,98]	38 [17,56]	64 [54,76]	44 [38,51]	92 [91,94]	47 [40,52]	69 [66,73]
rad. group B	27 [10,48]	96 [93,99]	35 [13,56]	62 [53,73]	43 [37,50]	93 [91,94]	46 [39,51]	69 [65,72]
rad. group C	23 [6,42]	99 [98,100]	36 [10,58]	61 [53,71]	31 [25,36]	99 [98,99]	42 [34,48]	65 [62,68]
radiologist average	27 [17,39]	97 [95,98]	36 [24,48]	62 [57,69]	39 [35,43]	95 [94,95]	46 [42,49]	68 [66,70]
R-Super†	32 [12,52]	97 [94,99]	41 [19,61]	89 [82,94]	53 [46,59]	95 [93,96]	57 [51,62]	88 [85,90]

Table D44: **Reader Study: tumor detection (R-Super trained without masks).** This table repeats the tumor detection comparison (malignant + benign vs no-tumor) using R-Super trained with CT-Report pairs only (no tumor masks). Even without masks, R-Super surpassed the radiologist average in sensitivity for 5 tumor types, achieving +7% sensitivity on average. These results demonstrate that report-only training is sufficient to surpass human performance in tumor detection for multiple tumor types. Blue indicates statistical significance ($p < 0.05$), comparing the AI model with radiologist average. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	36 [26,46]	98 [95,100]	52 [40,61]	67 [62,72]	55 [44,66]	97 [94,99]	69 [58,77]	76 [71,82]
rad. group B	37 [28,46]	98 [95,100]	53 [42,62]	68 [63,72]	55 [44,65]	97 [94,99]	69 [59,77]	76 [71,82]
rad. group C	23 [15,32]	100 [100,100]	37 [26,48]	61 [57,66]	4 [0,8]	100 [100,100]	7 [0,16]	52 [50,54]
radiologist average	32 [27,37]	98 [97,100]	48 [41,53]	65 [63,68]	38 [32,44]	98 [97,99]	54 [47,60]	68 [65,71]
R-Super	41 [31,50]	98 [95,100]	57 [46,66]	84 [78,89]	62 [51,72]	98 [95,100]	75 [66,82]	86 [80,91]

	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	47 [37,58]	95 [90,98]	60 [50,70]	72 [65,77]	74 [66,82]	85 [79,91]	77 [70,83]	82 [77,87]
rad. group B	49 [38,59]	95 [91,98]	62 [52,71]	72 [66,78]	75 [66,83]	84 [78,90]	77 [70,83]	82 [77,87]
rad. group C	24 [15,34]	100 [100,100]	39 [26,50]	62 [57,67]	69 [61,78]	98 [96,100]	81 [74,87]	84 [80,88]
radiologist average	40 [34,46]	96 [95,98]	55 [48,61]	69 [66,72]	73 [68,78]	89 [86,92]	78 [74,82]	83 [80,86]
R-Super	25 [16,34]	96 [92,99]	38 [27,49]	73 [66,80]	65 [56,73]	89 [83,94]	72 [65,79]	87 [82,91]

	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	38 [27,48]	96 [92,99]	52 [40,62]	67 [61,72]	8 [4,13]	100 [100,100]	14 [7,23]	54 [52,56]
rad. group B	38 [27,49]	96 [92,99]	52 [40,63]	67 [62,73]	5 [2,8]	100 [100,100]	9 [3,16]	52 [51,54]
rad. group C	14 [7,23]	100 [100,100]	25 [14,37]	57 [54,61]	48 [40,56]	98 [95,100]	64 [56,71]	73 [68,77]
radiologist average	30 [24,36]	97 [95,98]	44 [37,51]	64 [61,66]	20 [17,24]	99 [98,100]	34 [28,39]	60 [58,62]
R-Super	43 [32,53]	97 [94,99]	58 [47,68]	88 [82,92]	44 [35,53]	98 [96,100]	60 [52,69]	92 [88,95]

	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	43 [33,52]	96 [93,99]	58 [49,67]	70 [65,75]	43 [39,46]	95 [94,96]	55 [51,58]	70 [68,72]
rad. group B	43 [33,52]	97 [94,99]	59 [49,67]	70 [65,75]	43 [39,46]	95 [94,96]	54 [51,57]	70 [68,72]
rad. group C	35 [26,44]	100 [100,100]	52 [41,61]	67 [63,72]	31 [28,34]	99 [99,100]	44 [40,47]	65 [64,67]
radiologist average	40 [35,46]	98 [96,99]	56 [51,62]	69 [66,72]	39 [37,41]	96 [96,97]	53 [50,55]	68 [67,69]
R-Super	43 [33,52]	98 [95,100]	59 [49,68]	82 [76,88]	46 [43,50]	96 [95,98]	60 [56,64]	84 [82,86]

Table D45: **Reader Study: malignant tumor detection (R-Super trained without masks).** This table repeats the malignant tumor detection comparison (malignant vs no-tumor) using R-Super trained with CT-Report pairs only (no tumor masks). Even without masks, R-Super surpassed the radiologist average in sensitivity for 4 tumor types, and matched the radiologists on average (no significant difference in sensitivity). Unlike sensitivity, R-Super AUC significantly surpassed the radiologists on average (+16%), indicating that R-Super is operating at an unfavorable operating point at the radiologists' specificity but has stronger overall discriminative ability. Blue indicates statistical significance ($p < 0.05$), comparing the AI model with radiologist average. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder malignant tumor				esophagus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	57 [40,75]	98 [95,100]	69 [53,82]	78 [69,87]	67 [53,81]	97 [94,99]	76 [64,86]	83 [76,90]
rad. group B	57 [40,74]	98 [95,100]	69 [53,82]	78 [69,86]	67 [52,81]	97 [94,99]	76 [64,86]	83 [75,90]
rad. group C	37 [22,55]	100 [100,100]	54 [36,71]	69 [61,77]	2 [0,8]	100 [100,100]	4 [0,15]	51 [50,54]
radiologist average	50 [41,60]	98 [97,100]	65 [55,72]	75 [70,80]	46 [37,55]	98 [97,99]	60 [52,68]	72 [68,77]
R-Super	57 [40,74]	98 [95,100]	69 [54,82]	90 [81,97]	74 [60,86]	98 [95,100]	82 [71,91]	94 [90,98]

	gallbladder malignant tumor				uterus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	68 [48,87]	95 [90,98]	69 [51,82]	82 [72,92]	76 [61,89]	85 [79,91]	64 [50,75]	84 [75,91]
rad. group B	68 [50,85]	95 [91,98]	69 [53,82]	82 [72,92]	76 [61,89]	84 [78,90]	63 [50,75]	83 [75,91]
rad. group C	32 [13,52]	100 [100,100]	48 [23,68]	66 [56,76]	79 [64,92]	98 [96,100]	85 [74,94]	89 [82,96]
radiologist average	56 [44,68]	96 [95,98]	64 [55,74]	77 [71,83]	77 [68,84]	89 [86,92]	70 [62,76]	85 [80,89]
R-Super	48 [27,67]	96 [92,99]	57 [36,73]	89 [82,95]	82 [67,94]	89 [83,94]	72 [59,83]	92 [86,97]

	prostate malignant tumor				adrenal malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	42 [23,61]	96 [92,99]	51 [31,68]	69 [60,79]	13 [3,24]	100 [100,100]	23 [6,39]	56 [52,62]
rad. group B	42 [26,63]	96 [92,99]	51 [33,68]	69 [61,80]	8 [0,16]	100 [100,100]	14 [0,28]	54 [50,58]
rad. group C	27 [10,45]	100 [100,100]	42 [18,62]	64 [55,73]	36 [21,51]	98 [95,100]	50 [32,65]	67 [59,75]
radiologist average	37 [27,48]	97 [95,98]	49 [38,59]	67 [62,73]	19 [12,26]	99 [98,100]	31 [21,41]	59 [56,62]
R-Super	19 [5,36]	97 [94,99]	29 [8,49]	89 [81,95]	56 [40,71]	98 [96,100]	70 [55,82]	94 [89,98]

	spleen malignant tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	54 [32,77]	96 [93,99]	62 [41,78]	76 [64,88]	54 [48,60]	95 [94,96]	59 [53,64]	76 [72,79]
rad. group B	54 [32,76]	97 [94,99]	63 [42,79]	76 [65,87]	53 [47,59]	95 [94,96]	58 [52,63]	75 [72,78]
rad. group C	41 [21,64]	100 [100,100]	58 [34,78]	70 [60,82]	36 [30,43]	99 [99,100]	49 [41,55]	68 [65,71]
radiologist average	50 [38,61]	98 [96,99]	61 [50,71]	74 [68,80]	48 [44,52]	96 [96,97]	57 [53,61]	73 [71,75]
R-Super	32 [13,52]	98 [95,100]	44 [21,64]	81 [66,92]	53 [47,59]	96 [95,98]	60 [54,66]	90 [87,92]

Table D46: **Reader Study: malignancy identification (R-Super trained without masks).** This table repeats the malignancy identification comparison using R-Super trained with CT-Report pairs only (no tumor masks). Even without masks, R-Super surpassed the radiologist average in sensitivity for 4 tumor types, achieving +3% sensitivity on average. Blue indicates statistical significance ($p < 0.05$), comparing the AI model with radiologist average. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder malignancy id.				esophagus malignancy id.			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	46 [30,64]	97 [95,99]	58 [41,73]	72 [64,81]	63 [48,77]	93 [88,96]	67 [54,77]	79 [71,86]
rad. group B	43 [27,61]	97 [95,99]	56 [38,71]	70 [62,79]	65 [50,79]	93 [88,97]	68 [56,80]	80 [72,87]
rad. group C	37 [22,55]	99 [98,100]	53 [35,69]	68 [60,77]	2 [0,8]	99 [98,100]	4 [0,14]	51 [49,54]
radiologist average	42 [32,52]	98 [97,99]	56 [46,65]	70 [65,75]	43 [35,51]	95 [93,97]	54 [46,62]	70 [65,74]
R-Super	14 [3,26]	99 [97,100]	24 [6,39]	87 [77,94]	74 [60,86]	95 [92,99]	78 [67,87]	92 [87,96]

	gallbladder malignancy id.				uterus malignancy id.			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	56 [36,76]	93 [88,96]	56 [38,71]	76 [65,87]	64 [46,79]	76 [69,81]	42 [29,54]	72 [63,82]
rad. group B	56 [35,76]	94 [90,97]	58 [39,74]	76 [65,86]	64 [47,79]	77 [71,82]	42 [30,53]	73 [63,82]
rad. group C	28 [11,48]	100 [100,100]	44 [19,64]	64 [55,74]	73 [57,87]	97 [94,99]	76 [63,87]	85 [77,93]
radiologist average	47 [36,57]	96 [94,97]	54 [43,63]	72 [66,78]	67 [57,76]	83 [80,86]	50 [43,57]	77 [72,82]
R-Super	52 [32,73]	96 [92,99]	59 [38,76]	92 [86,96]	76 [60,90]	84 [79,89]	56 [42,67]	84 [76,92]

	prostate malignancy id.				adrenal malignancy id.			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	38 [20,58]	94 [90,98]	46 [25,63]	67 [57,77]	13 [3,24]	98 [96,100]	21 [6,36]	55 [50,61]
rad. group B	38 [21,59]	94 [90,98]	46 [26,63]	67 [58,78]	8 [0,16]	98 [96,100]	13 [0,27]	53 [49,57]
rad. group C	23 [8,39]	100 [100,100]	38 [14,56]	62 [54,70]	28 [14,43]	97 [94,99]	40 [23,56]	63 [56,71]
radiologist average	33 [23,43]	96 [94,98]	43 [31,54]	65 [60,70]	16 [10,23]	98 [96,99]	26 [17,35]	57 [54,61]
R-Super	15 [4,31]	96 [93,99]	23 [6,41]	61 [47,73]	36 [21,51]	98 [95,100]	49 [31,64]	83 [74,91]

	spleen malignancy id.				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	32 [13,55]	95 [92,98]	38 [17,56]	64 [54,76]	44 [38,51]	92 [91,94]	47 [40,52]	69 [66,73]
rad. group B	27 [10,48]	96 [93,99]	35 [13,56]	62 [53,73]	43 [37,50]	93 [91,94]	46 [39,51]	69 [65,72]
rad. group C	23 [6,42]	99 [98,100]	36 [10,58]	61 [53,71]	31 [25,36]	99 [98,99]	42 [34,48]	65 [62,68]
radiologist average	27 [17,39]	97 [95,98]	36 [24,48]	62 [57,69]	39 [35,43]	95 [94,95]	46 [42,49]	68 [66,70]
R-Super	23 [5,41]	97 [94,99]	31 [8,50]	79 [65,89]	42 [35,47]	95 [94,96]	46 [39,52]	82 [79,86]

Table D47: **Reader Study: Unlike R-Super, previous public AI models do not surpass human performance in tumor detection.** In tumor detection (benign + malignant vs no-tumor), we compared the radiologist performance to ULS (segmentation) and Merlin (zero-shot classification), on the reader study test dataset. These are the previous public models that, after R-Super, achieved the highest performance in our experiments (Fig. 2). Unlike R-Super, these previous public AI models were not able to match or surpass the average radiologist sensitivity in any tumor type, at matched specificity. Blue indicates statistical significance ($p < 0.05$), comparing the best previous public model with radiologist average. R-Super (ours, included in the table for reference) was excluded from the comparison. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	36 [26,46]	98 [95,100]	52 [40,61]	67 [62,72]	55 [44,66]	97 [94,99]	69 [58,77]	76 [71,82]
rad. group B	37 [28,46]	98 [95,100]	53 [42,62]	68 [63,72]	55 [44,65]	97 [94,99]	69 [59,77]	76 [71,82]
rad. group C	23 [15,32]	100 [100,100]	37 [26,48]	61 [57,66]	4 [0,8]	100 [100,100]	7 [0,16]	52 [50,54]
radiologist average	32 [27,37]	98 [97,100]	48 [41,53]	65 [63,68]	38 [32,44]	98 [97,99]	54 [47,60]	68 [65,71]
ULS	12 [5,20]	92 [85,98]	20 [9,31]	54 [45,64]	11 [3,19]	98 [95,100]	19 [6,31]	53 [46,60]
Merlin (zero-shot)	13 [6,21]	98 [95,100]	22 [11,34]	54 [44,63]	6 [1,12]	98 [95,100]	11 [3,22]	56 [46,66]
R-Super†	48 [38,58]	98 [95,100]	64 [54,72]	88 [82,92]	67 [57,77]	98 [95,100]	79 [70,86]	92 [87,96]

	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	47 [37,58]	95 [90,98]	60 [50,70]	72 [65,77]	74 [66,82]	85 [79,91]	77 [70,83]	82 [77,87]
rad. group B	49 [38,59]	95 [91,98]	62 [52,71]	72 [66,78]	75 [66,83]	84 [78,90]	77 [70,83]	82 [77,87]
rad. group C	24 [15,34]	100 [100,100]	39 [26,50]	62 [57,67]	69 [61,78]	98 [96,100]	81 [74,87]	84 [80,88]
radiologist average	40 [34,46]	96 [95,98]	55 [48,61]	69 [66,72]	73 [68,78]	89 [86,92]	78 [74,82]	83 [80,86]
ULS	10 [2,19]	97 [92,100]	18 [4,30]	58 [52,64]	15 [8,22]	89 [82,96]	24 [13,35]	54 [46,62]
Merlin (zero-shot)	4 [0,10]	97 [92,100]	8 [0,18]	51 [40,61]	26 [17,36]	89 [81,97]	39 [28,50]	63 [55,72]
R-Super†	69 [59,78]	96 [93,99]	79 [71,85]	90 [85,94]	63 [54,72]	89 [83,94]	71 [64,78]	82 [77,87]

	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	38 [27,48]	96 [92,99]	52 [40,62]	67 [61,72]	8 [4,13]	100 [100,100]	14 [7,23]	54 [52,56]
rad. group B	38 [27,49]	96 [92,99]	52 [40,63]	67 [62,73]	5 [2,8]	100 [100,100]	9 [3,16]	52 [51,54]
rad. group C	14 [7,23]	100 [100,100]	25 [14,37]	57 [54,61]	48 [40,56]	98 [95,100]	64 [56,71]	73 [68,77]
radiologist average	30 [24,36]	97 [95,98]	44 [37,51]	64 [61,66]	20 [17,24]	99 [98,100]	34 [28,39]	60 [58,62]
ULS	21 [8,37]	97 [92,100]	33 [14,52]	60 [49,71]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Merlin (zero-shot)	9 [0,19]	97 [92,100]	16 [0,30]	63 [50,75]	5 [1,9]	100 [100,100]	9 [2,16]	68 [59,76]
R-Super†	54 [43,66]	97 [94,99]	68 [58,77]	90 [85,95]	54 [46,63]	99 [98,100]	70 [62,77]	92 [89,95]

	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	43 [33,52]	96 [93,99]	58 [49,67]	70 [65,75]	43 [39,46]	95 [94,96]	55 [51,58]	70 [68,72]
rad. group B	43 [33,52]	97 [94,99]	59 [49,67]	70 [65,75]	43 [39,46]	95 [94,96]	54 [51,57]	70 [68,72]
rad. group C	35 [26,44]	100 [100,100]	52 [41,61]	67 [63,72]	31 [28,34]	99 [99,100]	44 [40,47]	65 [64,67]
radiologist average	40 [35,46]	98 [96,99]	56 [51,62]	69 [66,72]	39 [37,41]	96 [96,97]	53 [50,55]	68 [67,69]
ULS	0 [0,0]	98 [95,100]	0 [0,0]	49 [41,58]	10 [7,13]	96 [94,98]	16 [12,20]	54 [51,57]
Merlin (zero-shot)	10 [4,16]	98 [95,100]	17 [7,28]	60 [50,69]	10 [7,13]	97 [94,99]	17 [13,22]	59 [53,65]
R-Super†	71 [62,79]	98 [95,100]	82 [75,87]	86 [80,91]	61 [57,65]	96 [95,98]	73 [70,76]	88 [87,90]

Table D48: **Reader Study: Unlike R-Super, previous public AI models do not surpass human performance in malignant tumor detection.** In malignant tumor detection (malignant vs no-tumor), we compared the radiologist performance to ULS (segmentation) and Merlin (zero-shot classification), on the reader study test dataset. These are the previous public models that, after R-Super, achieved the highest performance in our experiments (Fig. 2). Unlike R-Super, these previous public AI models were not able to match or surpass the average radiologist sensitivity in any tumor type, at matched specificity. Blue indicates statistical significance ($p < 0.05$), comparing the best previous public model with radiologist average. R-Super (ours, included in the table for reference) was excluded from the comparison. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder malignant tumor				esophagus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	57 [40,75]	98 [95,100]	69 [53,82]	78 [69,87]	67 [53,81]	97 [94,99]	76 [64,86]	83 [76,90]
rad. group B	57 [40,74]	98 [95,100]	69 [53,82]	78 [69,86]	67 [52,81]	97 [94,99]	76 [64,86]	83 [75,90]
rad. group C	37 [22,55]	100 [100,100]	54 [36,71]	69 [61,77]	2 [0,8]	100 [100,100]	4 [0,15]	51 [50,54]
radiologist average	50 [41,60]	98 [97,100]	65 [55,72]	75 [70,80]	46 [37,55]	98 [97,99]	60 [52,68]	72 [68,77]
ULS	11 [2,23]	92 [85,98]	18 [4,34]	56 [45,68]	14 [4,26]	98 [95,100]	24 [8,39]	55 [48,63]
Merlin (zero-shot)	23 [10,37]	98 [95,100]	36 [18,53]	59 [47,72]	2 [0,8]	98 [95,100]	4 [0,15]	58 [47,69]
R-Super†	71 [56,85]	98 [95,100]	79 [67,89]	95 [89,99]	72 [58,85]	98 [95,100]	80 [69,89]	96 [93,98]

	gallbladder malignant tumor				uterus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	68 [48,87]	95 [90,98]	69 [51,82]	82 [72,92]	76 [61,89]	85 [79,91]	64 [50,75]	84 [75,91]
rad. group B	68 [50,85]	95 [91,98]	69 [53,82]	82 [72,92]	76 [61,89]	84 [78,90]	63 [50,75]	83 [75,91]
rad. group C	32 [13,52]	100 [100,100]	48 [23,68]	66 [56,76]	79 [64,92]	98 [96,100]	85 [74,94]	89 [82,96]
radiologist average	56 [44,68]	96 [95,98]	64 [55,74]	77 [71,83]	77 [68,84]	89 [86,92]	70 [62,76]	85 [80,89]
ULS	16 [4,31]	97 [92,100]	26 [6,45]	62 [53,72]	12 [2,25]	89 [82,96]	18 [4,34]	58 [48,69]
Merlin (zero-shot)	4 [0,13]	97 [92,100]	7 [0,22]	50 [37,63]	36 [20,52]	89 [81,97]	46 [28,61]	66 [53,80]
R-Super†	96 [88,100]	96 [93,99]	89 [79,96]	99 [98,100]	79 [64,92]	89 [83,94]	70 [58,81]	86 [78,93]

	prostate malignant tumor				adrenal malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	42 [23,61]	96 [92,99]	51 [31,68]	69 [60,79]	13 [3,24]	100 [100,100]	23 [6,39]	56 [52,62]
rad. group B	42 [26,63]	96 [92,99]	51 [33,68]	69 [61,80]	8 [0,16]	100 [100,100]	14 [0,28]	54 [50,58]
rad. group C	27 [10,45]	100 [100,100]	42 [18,62]	64 [55,73]	36 [21,51]	98 [95,100]	50 [32,65]	67 [59,75]
radiologist average	37 [27,48]	97 [95,98]	49 [38,59]	67 [62,73]	19 [12,26]	99 [98,100]	31 [21,41]	59 [56,62]
ULS	12 [0,26]	97 [92,100]	19 [0,38]	54 [44,66]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Merlin (zero-shot)	12 [0,25]	97 [92,100]	19 [0,37]	68 [53,82]	5 [0,13]	100 [100,100]	10 [0,23]	73 [62,83]
R-Super†	65 [46,83]	97 [94,99]	72 [56,85]	92 [83,98]	67 [50,81]	99 [98,100]	79 [66,88]	94 [89,99]

	spleen malignant tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	54 [32,77]	96 [93,99]	62 [41,78]	76 [64,88]	54 [48,60]	95 [94,96]	59 [53,64]	76 [72,79]
rad. group B	54 [32,76]	97 [94,99]	63 [42,79]	76 [65,87]	53 [47,59]	95 [94,96]	58 [52,63]	75 [72,78]
rad. group C	41 [21,64]	100 [100,100]	58 [34,78]	70 [60,82]	36 [30,43]	99 [99,100]	49 [41,55]	68 [65,71]
radiologist average	50 [38,61]	98 [96,99]	61 [50,71]	74 [68,80]	48 [44,52]	96 [96,97]	57 [53,61]	73 [71,75]
ULS	0 [0,0]	98 [95,100]	0 [0,0]	58 [46,71]	9 [6,13]	96 [94,98]	15 [10,20]	56 [52,60]
Merlin (zero-shot)	14 [0,30]	98 [95,100]	23 [0,44]	68 [53,81]	14 [9,18]	97 [94,99]	21 [14,27]	63 [56,70]
R-Super†	77 [58,94]	98 [95,100]	81 [67,92]	94 [88,99]	75 [69,81]	96 [95,98]	79 [74,83]	94 [92,96]

D.8.1 Reader Study (High Confidence)

Here, we compare R-Super to radiologists at the certain confidence level. As described in Section 2.4, for each CT scan, radiologists indicated which organs probably or certainly contained a tumor, and which probably or certainly contained a malignant tumor. At the certain level, a case is counted as positive only if the radiologist indicated certainty that a tumor was present or malignant; cases marked as only probable are not counted as positives. This is a stricter comparison than the probable level, as radiologists are less sensitive when required to express full confidence. In the tables below, R-Super is evaluated at specificity matched to the radiologists'.

Table D49: **Reader Study (certain level): tumor detection.** R-Super significantly (p<0.05) surpassed the radiologist average in sensitivity for 4 tumor types and achieved +18% sensitivity and +20% AUC on average. Blue indicates statistical significance ($p < 0.05$), comparing the AI model with radiologist average. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	26 [18,35]	100 [100,100]	41 [30,52]	67 [62,72]	40 [29,50]	100 [100,100]	57 [45,67]	76 [71,82]
rad. group B	27 [18,36]	100 [100,100]	42 [30,53]	68 [63,72]	40 [30,51]	100 [100,100]	57 [46,68]	76 [71,82]
rad. group C	16 [10,24]	100 [100,100]	28 [18,39]	61 [57,66]	2 [0,6]	100 [100,100]	5 [0,12]	52 [50,54]
radiologist average	23 [18,28]	100 [100,100]	37 [31,43]	65 [63,68]	28 [22,34]	100 [100,100]	43 [36,50]	68 [65,71]
R-Super†	22 [14,30]	100 [100,100]	36 [24,46]	88 [82,92]	65 [53,75]	100 [100,100]	78 [70,86]	92 [87,96]

	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	37 [28,48]	98 [96,100]	53 [42,64]	72 [65,77]	66 [57,74]	93 [89,97]	76 [69,82]	82 [77,87]
rad. group B	38 [27,48]	98 [96,100]	54 [42,64]	72 [66,78]	66 [57,75]	93 [89,97]	76 [69,82]	82 [77,87]
rad. group C	22 [13,32]	100 [100,100]	36 [23,48]	62 [57,67]	66 [56,74]	99 [98,100]	79 [72,85]	84 [80,88]
radiologist average	32 [27,38]	99 [98,100]	48 [42,55]	69 [66,72]	66 [61,71]	95 [93,97]	77 [73,80]	83 [80,86]
R-Super†	49 [38,60]	99 [97,100]	66 [55,75]	90 [85,94]	35 [26,44]	95 [91,98]	50 [39,59]	82 [77,87]

	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	25 [16,34]	98 [96,100]	39 [27,50]	67 [61,72]	0 [0,0]	100 [100,100]	0 [0,0]	54 [52,56]
rad. group B	25 [16,35]	98 [96,100]	39 [27,51]	67 [62,73]	1 [0,2]	100 [100,100]	2 [0,5]	52 [51,54]
rad. group C	10 [4,18]	100 [100,100]	19 [8,30]	57 [54,61]	47 [38,55]	98 [95,100]	63 [54,70]	73 [68,77]
radiologist average	20 [15,25]	99 [98,100]	33 [26,40]	64 [61,67]	16 [12,20]	99 [98,100]	27 [22,32]	60 [58,62]
R-Super†	54 [43,66]	98 [95,100]	69 [59,78]	90 [85,95]	54 [46,63]	99 [98,100]	70 [62,77]	92 [89,95]

	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	41 [32,50]	98 [96,100]	58 [48,66]	70 [65,75]	34 [30,37]	98 [98,99]	46 [42,50]	70 [68,72]
rad. group B	41 [32,50]	98 [96,100]	58 [48,66]	70 [65,75]	34 [31,37]	98 [98,99]	47 [43,50]	70 [68,72]
rad. group C	34 [25,42]	100 [100,100]	51 [40,60]	67 [63,72]	28 [25,31]	100 [99,100]	40 [36,44]	65 [64,67]
radiologist average	39 [34,44]	99 [98,100]	56 [50,61]	69 [66,72]	32 [30,34]	99 [98,99]	46 [44,48]	68 [67,69]
R-Super†	67 [58,76]	99 [98,100]	80 [73,86]	86 [80,91]	50 [46,53]	99 [98,99]	64 [60,67]	88 [87,90]

Table D50: **Reader Study (certain level): malignant tumor detection.** R-Super surpassed the radiologist average in sensitivity for 6 tumor types, with +25% average sensitivity at approximately matched specificity. Blue indicates statistical significance ($p < 0.05$), comparing the AI model with radiologist average. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder malignant tumor				esophagus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	46 [30,62]	100 [100,100]	63 [46,77]	78 [69,87]	54 [39,68]	100 [100,100]	70 [56,81]	83 [76,90]
rad. group B	46 [29,62]	100 [100,100]	63 [45,76]	78 [69,86]	54 [38,69]	100 [100,100]	70 [55,82]	83 [75,90]
rad. group C	23 [10,38]	100 [100,100]	37 [19,55]	69 [61,77]	2 [0,8]	100 [100,100]	4 [0,15]	51 [50,54]
radiologist average	38 [29,47]	100 [100,100]	55 [45,64]	75 [70,80]	36 [28,45]	100 [100,100]	53 [44,62]	72 [68,77]
R-Super†	43 [27,59]	100 [100,100]	60 [42,74]	95 [89,99]	67 [53,82]	100 [100,100]	81 [69,90]	96 [93,98]

	gallbladder malignant tumor				uterus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	60 [40,79]	98 [96,100]	71 [53,86]	82 [72,92]	76 [61,89]	93 [89,97]	75 [61,85]	84 [75,91]
rad. group B	60 [39,80]	98 [96,100]	71 [52,86]	82 [72,92]	76 [61,89]	93 [89,97]	75 [62,84]	83 [75,91]
rad. group C	24 [8,41]	100 [100,100]	39 [14,58]	66 [56,76]	70 [54,85]	99 [98,100]	81 [68,91]	89 [82,96]
radiologist average	48 [37,60]	99 [98,100]	63 [52,72]	77 [71,83]	74 [64,82]	95 [93,97]	76 [68,83]	85 [81,90]
R-Super†	88 [73,100]	99 [97,100]	92 [81,98]	99 [98,100]	64 [47,80]	95 [91,98]	69 [55,81]	86 [78,93]

	prostate malignant tumor				adrenal malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	35 [17,53]	98 [96,100]	49 [27,67]	69 [60,79]	0 [0,0]	100 [100,100]	0 [0,0]	56 [52,62]
rad. group B	35 [18,54]	98 [96,100]	49 [28,67]	69 [61,80]	0 [0,0]	100 [100,100]	0 [0,0]	54 [50,58]
rad. group C	23 [8,40]	100 [100,100]	38 [15,57]	64 [55,73]	36 [21,51]	98 [95,100]	50 [32,65]	67 [59,75]
radiologist average	31 [21,42]	99 [98,100]	45 [33,57]	67 [62,73]	12 [6,19]	99 [98,100]	21 [12,31]	59 [56,63]
R-Super†	65 [46,83]	98 [95,100]	74 [58,86]	92 [83,98]	67 [50,81]	99 [98,100]	79 [66,88]	94 [89,99]

	spleen malignant tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	54 [32,77]	98 [96,100]	67 [45,83]	76 [64,88]	46 [40,53]	98 [98,99]	56 [50,61]	76 [72,79]
rad. group B	54 [32,76]	98 [96,100]	67 [45,82]	76 [65,87]	46 [40,52]	98 [98,99]	56 [50,61]	75 [72,78]
rad. group C	41 [21,64]	100 [100,100]	58 [34,78]	70 [60,82]	31 [25,38]	100 [99,100]	44 [36,50]	68 [65,71]
radiologist average	50 [38,61]	99 [98,100]	64 [52,74]	74 [68,81]	41 [37,45]	99 [98,99]	54 [50,58]	73 [71,75]
R-Super†	68 [46,87]	99 [98,100]	79 [61,91]	94 [88,99]	66 [59,72]	99 [98,99]	76 [71,81]	94 [92,96]

Table D51: **Reader Study (certain level): malignancy identification.** At the certain level, R-Super surpassed the radiologist average in sensitivity for 3 tumor types in malignancy identification, and it had a similar sensitivity on average. The R-Super's average sensitivity is reduced by tumor types where the radiologist achieved 100% specificity, but the model could not perform well at this extreme specificity, even if R-Super's AUC was considerably higher for these tumor types (gallbladder and spleen). Blue indicates statistical significance ($p < 0.05$), comparing the AI model with radiologist average. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder malignancy id.				esophagus malignancy id.			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	29 [14,46]	99 [98,100]	44 [24,61]	72 [64,81]	56 [40,71]	97 [94,99]	67 [53,78]	79 [71,86]
rad. group B	29 [14,45]	99 [97,100]	43 [24,60]	70 [62,79]	56 [40,70]	97 [93,99]	67 [53,78]	80 [72,87]
rad. group C	17 [6,32]	100 [100,100]	29 [11,49]	68 [60,77]	2 [0,8]	100 [100,100]	4 [0,15]	51 [49,54]
radiologist average	25 [16,33]	99 [98,100]	39 [28,49]	70 [65,75]	38 [30,46]	98 [96,99]	52 [43,60]	70 [65,74]
R-Super†	57 [40,73]	99 [97,100]	70 [55,82]	93 [86,98]	58 [43,72]	98 [95,100]	70 [56,81]	90 [85,95]

	gallbladder malignancy id.				uterus malignancy id.			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	48 [28,68]	99 [98,100]	63 [42,79]	76 [65,87]	39 [24,56]	94 [90,97]	45 [29,61]	72 [63,82]
rad. group B	48 [28,69]	99 [98,100]	63 [42,80]	76 [65,86]	39 [24,56]	94 [90,97]	45 [29,59]	73 [63,82]
rad. group C	16 [3,32]	100 [100,100]	28 [7,48]	64 [55,74]	39 [23,56]	100 [98,100]	55 [36,70]	85 [77,93]
radiologist average	37 [26,50]	100 [99,100]	53 [40,65]	72 [66,78]	39 [30,49]	96 [94,97]	48 [39,57]	77 [71,82]
R-Super†	4 [0,14]	100 [100,100]	8 [0,24]	95 [90,99]	33 [18,49]	96 [93,98]	42 [25,57]	77 [66,86]

	prostate malignancy id.				adrenal malignancy id.			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	27 [11,44]	100 [100,100]	42 [20,62]	67 [57,77]	0 [0,0]	100 [100,100]	0 [0,0]	55 [50,61]
rad. group B	27 [11,46]	100 [100,100]	42 [20,63]	67 [58,78]	0 [0,0]	100 [100,100]	0 [0,0]	53 [49,57]
rad. group C	15 [3,30]	100 [100,100]	27 [5,46]	62 [54,70]	18 [6,31]	100 [98,100]	30 [12,46]	63 [56,71]
radiologist average	23 [13,32]	100 [100,100]	38 [24,50]	65 [60,71]	6 [2,11]	100 [99,100]	11 [5,19]	57 [54,61]
R-Super†	19 [5,36]	100 [100,100]	32 [10,53]	91 [83,97]	8 [0,17]	100 [100,100]	14 [0,29]	79 [69,88]

	spleen malignancy id.				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	18 [4,36]	100 [100,100]	31 [8,53]	64 [54,76]	31 [25,37]	98 [98,99]	42 [34,48]	69 [66,73]
rad. group B	18 [4,36]	100 [100,100]	31 [8,53]	62 [53,73]	31 [25,37]	98 [98,99]	42 [34,48]	69 [65,72]
rad. group C	23 [6,42]	100 [100,100]	37 [11,59]	61 [53,71]	19 [13,24]	100 [100,100]	30 [22,37]	65 [62,68]
radiologist average	20 [10,29]	100 [100,100]	33 [19,46]	62 [57,68]	27 [24,30]	99 [98,99]	39 [35,43]	68 [66,70]
R-Super†	14 [0,30]	100 [100,100]	24 [0,47]	89 [82,94]	28 [23,33]	99 [98,99]	37 [31,43]	88 [85,90]

Table D52: **Reader Study (certain level): tumor detection (R-Super trained without masks).** At the certain level, R-Super without masks surpassed the radiologist average in sensitivity for 3 tumor types, and achieved a similar sensitivity on average. For several tumor types, R-Super had lower sensitivity than radiologists but achieved higher AUC, indicating that the model has stronger overall discriminative ability but operates at an unfavorable operating point at the radiologists' 'certain' specificity level. Blue indicates statistical significance ($p < 0.05$), comparing the AI model with radiologist average. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	26 [18,35]	100 [100,100]	41 [30,52]	67 [62,72]	40 [29,50]	100 [100,100]	57 [45,67]	76 [71,82]
rad. group B	27 [18,36]	100 [100,100]	42 [30,53]	68 [63,72]	40 [30,51]	100 [100,100]	57 [46,68]	76 [71,82]
rad. group C	16 [10,24]	100 [100,100]	28 [18,39]	61 [57,66]	2 [0,6]	100 [100,100]	5 [0,12]	52 [50,54]
radiologist average	23 [18,28]	100 [100,100]	37 [31,43]	65 [63,68]	28 [22,34]	100 [100,100]	43 [36,50]	68 [65,71]
R-Super	14 [8,22]	100 [100,100]	25 [14,36]	84 [78,89]	51 [40,62]	100 [100,100]	68 [57,76]	86 [80,91]

	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	37 [28,48]	98 [96,100]	53 [42,64]	72 [65,77]	66 [57,74]	93 [89,97]	76 [69,82]	82 [77,87]
rad. group B	38 [27,48]	98 [96,100]	54 [42,64]	72 [66,78]	66 [57,75]	93 [89,97]	76 [69,82]	82 [77,87]
rad. group C	22 [13,32]	100 [100,100]	36 [23,48]	62 [57,67]	66 [56,74]	99 [98,100]	79 [72,85]	84 [80,88]
radiologist average	32 [27,38]	99 [98,100]	48 [42,55]	69 [66,72]	66 [61,71]	95 [93,97]	77 [73,80]	83 [80,86]
R-Super	14 [7,23]	99 [98,100]	25 [13,37]	73 [66,80]	47 [37,57]	95 [91,98]	61 [52,70]	87 [82,91]

	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	25 [16,34]	98 [96,100]	39 [27,50]	67 [61,72]	0 [0,0]	100 [100,100]	0 [0,0]	54 [52,56]
rad. group B	25 [16,35]	98 [96,100]	39 [27,51]	67 [62,73]	1 [0,2]	100 [100,100]	2 [0,5]	52 [51,54]
rad. group C	10 [4,18]	100 [100,100]	19 [8,30]	57 [54,61]	47 [38,55]	98 [95,100]	63 [54,70]	73 [68,77]
radiologist average	20 [15,25]	99 [98,100]	33 [26,40]	64 [61,67]	16 [12,20]	99 [98,100]	27 [22,32]	60 [58,62]
R-Super	31 [22,41]	99 [98,100]	47 [35,58]	88 [82,92]	44 [35,53]	98 [96,100]	60 [52,69]	92 [88,95]

	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	41 [32,50]	98 [96,100]	58 [48,66]	70 [65,75]	34 [30,37]	98 [98,99]	46 [42,50]	70 [68,72]
rad. group B	41 [32,50]	98 [96,100]	58 [48,66]	70 [65,75]	34 [31,37]	98 [98,99]	47 [43,50]	70 [68,72]
rad. group C	34 [25,42]	100 [100,100]	51 [40,60]	67 [63,72]	28 [25,31]	100 [99,100]	40 [36,44]	65 [64,67]
radiologist average	39 [34,44]	99 [98,100]	56 [50,61]	69 [66,72]	32 [30,34]	99 [98,99]	46 [44,48]	68 [67,69]
R-Super	43 [33,52]	98 [95,100]	59 [49,68]	82 [76,88]	35 [32,39]	98 [98,99]	49 [45,53]	84 [82,86]

Table D53: **Reader Study (certain level): malignant tumor detection (R-Super trained without masks).** R-Super without masks surpassed the radiologist average in sensitivity for 2 tumor types. For several tumor types, R-Super had lower sensitivity than radiologists but achieved higher AUC, indicating that the model has stronger overall discriminative ability but operates at an unfavorable operating point at the radiologists' 'certain' specificity level. Blue indicates statistical significance ($p < 0.05$), comparing the AI model with radiologist average. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder malignant tumor				esophagus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	46 [30,62]	100 [100,100]	63 [46,77]	78 [69,87]	54 [39,68]	100 [100,100]	70 [56,81]	83 [76,90]
rad. group B	46 [29,62]	100 [100,100]	63 [45,76]	78 [69,86]	54 [38,69]	100 [100,100]	70 [55,82]	83 [75,90]
rad. group C	23 [10,38]	100 [100,100]	37 [19,55]	69 [61,77]	2 [0,8]	100 [100,100]	4 [0,15]	51 [50,54]
radiologist average	38 [29,47]	100 [100,100]	55 [45,64]	75 [70,80]	36 [28,45]	100 [100,100]	53 [44,62]	72 [68,77]
R-Super	23 [10,36]	100 [100,100]	37 [18,53]	90 [81,97]	67 [53,81]	100 [100,100]	81 [70,90]	94 [90,98]

	gallbladder malignant tumor				uterus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	60 [40,79]	98 [96,100]	71 [53,86]	82 [72,92]	76 [61,89]	93 [89,97]	75 [61,85]	84 [75,91]
rad. group B	60 [39,80]	98 [96,100]	71 [52,86]	82 [72,92]	76 [61,89]	93 [89,97]	75 [62,84]	83 [75,91]
rad. group C	24 [8,41]	100 [100,100]	39 [14,58]	66 [56,76]	70 [54,85]	99 [98,100]	81 [68,91]	89 [82,96]
radiologist average	48 [37,60]	99 [98,100]	63 [52,72]	77 [71,83]	74 [64,82]	95 [93,97]	76 [68,83]	85 [81,90]
R-Super	32 [15,53]	99 [98,100]	47 [24,68]	89 [82,95]	70 [53,86]	95 [91,98]	73 [59,84]	92 [86,97]

	prostate malignant tumor				adrenal malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	35 [17,53]	98 [96,100]	49 [27,67]	69 [60,79]	0 [0,0]	100 [100,100]	0 [0,0]	56 [52,62]
rad. group B	35 [18,54]	98 [96,100]	49 [28,67]	69 [61,80]	0 [0,0]	100 [100,100]	0 [0,0]	54 [50,58]
rad. group C	23 [8,40]	100 [100,100]	38 [15,57]	64 [55,73]	36 [21,51]	98 [95,100]	50 [32,65]	67 [59,75]
radiologist average	31 [21,42]	99 [98,100]	45 [33,57]	67 [62,73]	12 [6,19]	99 [98,100]	21 [12,31]	59 [56,63]
R-Super	15 [4,31]	99 [98,100]	26 [7,46]	89 [81,95]	56 [40,71]	98 [96,100]	70 [55,82]	94 [89,98]

	spleen malignant tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	54 [32,77]	98 [96,100]	67 [45,83]	76 [64,88]	46 [40,53]	98 [98,99]	56 [50,61]	76 [72,79]
rad. group B	54 [32,76]	98 [96,100]	67 [45,82]	76 [65,87]	46 [40,52]	98 [98,99]	56 [50,61]	75 [72,78]
rad. group C	41 [21,64]	100 [100,100]	58 [34,78]	70 [60,82]	31 [25,38]	100 [99,100]	44 [36,50]	68 [65,71]
radiologist average	50 [38,61]	99 [98,100]	64 [52,74]	74 [68,81]	41 [37,45]	99 [98,99]	54 [50,58]	73 [71,75]
R-Super	32 [13,52]	98 [95,100]	44 [21,64]	81 [66,92]	42 [36,48]	98 [98,99]	54 [47,60]	90 [87,92]

Table D54: **Reader Study (certain level): malignancy identification (R-Super trained without masks).** Without masks, R-Super surpassed the radiologist average in sensitivity for 3 tumor types. For several tumor types, R-Super sensitivity collapsed to 0, because of the selected operating point of 100% specificity. However, for several of these types the R-Super AUC surpassed the radiologists, indicating that the model has stronger overall discriminative ability but operates at an extremely unfavorable operating point at the radiologists' 'certain' specificity level. Blue indicates statistical significance ($p < 0.05$), comparing the AI model with radiologist average. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder malignancy id.				esophagus malignancy id.			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	29 [14,46]	99 [98,100]	44 [24,61]	72 [64,81]	56 [40,71]	97 [94,99]	67 [53,78]	79 [71,86]
rad. group B	29 [14,45]	99 [97,100]	43 [24,60]	70 [62,79]	56 [40,70]	97 [93,99]	67 [53,78]	80 [72,87]
rad. group C	17 [6,32]	100 [100,100]	29 [11,49]	68 [60,77]	2 [0,8]	100 [100,100]	4 [0,15]	51 [49,54]
radiologist average	25 [16,33]	99 [98,100]	39 [28,49]	70 [65,75]	38 [30,46]	98 [96,99]	52 [43,60]	70 [65,74]
R-Super	14 [3,26]	99 [97,100]	24 [6,39]	87 [77,94]	63 [48,77]	98 [95,100]	74 [61,84]	92 [87,96]

	gallbladder malignancy id.				uterus malignancy id.			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	48 [28,68]	99 [98,100]	63 [42,79]	76 [65,87]	39 [24,56]	94 [90,97]	45 [29,61]	72 [63,82]
rad. group B	48 [28,69]	99 [98,100]	63 [42,80]	76 [65,86]	39 [24,56]	94 [90,97]	45 [29,59]	73 [63,82]
rad. group C	16 [3,32]	100 [100,100]	28 [7,48]	64 [55,74]	39 [23,56]	100 [98,100]	55 [36,70]	85 [77,93]
radiologist average	37 [26,50]	100 [99,100]	53 [40,65]	72 [66,78]	39 [30,49]	96 [94,97]	48 [39,57]	77 [71,82]
R-Super	0 [0,0]	100 [100,100]	0 [0,0]	92 [86,96]	42 [26,59]	96 [93,99]	52 [33,67]	84 [76,92]

	prostate malignancy id.				adrenal malignancy id.			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	27 [11,44]	100 [100,100]	42 [20,62]	67 [57,77]	0 [0,0]	100 [100,100]	0 [0,0]	55 [50,61]
rad. group B	27 [11,46]	100 [100,100]	42 [20,63]	67 [58,78]	0 [0,0]	100 [100,100]	0 [0,0]	53 [49,57]
rad. group C	15 [3,30]	100 [100,100]	27 [5,46]	62 [54,70]	18 [6,31]	100 [98,100]	30 [12,46]	63 [56,71]
radiologist average	23 [13,32]	100 [100,100]	38 [24,50]	65 [60,71]	6 [2,11]	100 [99,100]	11 [5,19]	57 [54,61]
R-Super	8 [0,19]	100 [100,100]	14 [0,32]	61 [47,73]	23 [10,36]	100 [100,100]	38 [19,53]	83 [74,91]

	spleen malignancy id.				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	18 [4,36]	100 [100,100]	31 [8,53]	64 [54,76]	31 [25,37]	98 [98,99]	42 [34,48]	69 [66,73]
rad. group B	18 [4,36]	100 [100,100]	31 [8,53]	62 [53,73]	31 [25,37]	98 [98,99]	42 [34,48]	69 [65,72]
rad. group C	23 [6,42]	100 [100,100]	37 [11,59]	61 [53,71]	19 [13,24]	100 [100,100]	30 [22,37]	65 [62,68]
radiologist average	20 [10,29]	100 [100,100]	33 [19,46]	62 [57,68]	27 [24,30]	99 [98,99]	39 [35,43]	68 [66,70]
R-Super	0 [0,0]	100 [100,100]	0 [0,0]	79 [65,89]	22 [17,26]	99 [98,100]	29 [23,33]	82 [79,86]

Table D55: **Reader Study (certain level): Unlike R-Super, previous public AI models do not surpass human performance in tumor detection.** In tumor detection (benign + malignant vs no-tumor), we compared the radiologist performance to ULS (segmentation) and Merlin (zero-shot classification), on the reader study test dataset. These are the previous public models that, after R-Super, achieved the highest performance in our experiments (Fig. 2). Unlike R-Super, these previous public AI models were not able to match or surpass the average radiologist sensitivity in any tumor type, at matched specificity. Blue indicates statistical significance ($p < 0.05$), comparing the best previous public model with radiologist average. R-Super (ours, included in the table for reference) was excluded from the comparison. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder tumor				esophagus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	26 [18,35]	100 [100,100]	41 [30,52]	67 [62,72]	40 [29,50]	100 [100,100]	57 [45,67]	76 [71,82]
rad. group B	27 [18,36]	100 [100,100]	42 [30,53]	68 [63,72]	40 [30,51]	100 [100,100]	57 [46,68]	76 [71,82]
rad. group C	16 [10,24]	100 [100,100]	28 [18,39]	61 [57,66]	2 [0,6]	100 [100,100]	5 [0,12]	52 [50,54]
radiologist average	23 [18,28]	100 [100,100]	37 [31,43]	65 [63,68]	28 [22,34]	100 [100,100]	43 [36,50]	68 [65,71]
ULS	12 [5,20]	92 [85,98]	20 [9,31]	54 [45,64]	9 [3,17]	100 [100,100]	17 [5,29]	53 [46,60]
Merlin (zero-shot)	6 [1,12]	100 [100,100]	12 [2,21]	54 [44,63]	2 [0,5]	100 [100,100]	3 [0,9]	56 [46,66]
R-Super†	22 [14,30]	100 [100,100]	36 [24,46]	88 [82,92]	65 [53,75]	100 [100,100]	78 [70,86]	92 [87,96]

	gallbladder tumor				uterus tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	37 [28,48]	98 [96,100]	53 [42,64]	72 [65,77]	66 [57,74]	93 [89,97]	76 [69,82]	82 [77,87]
rad. group B	38 [27,48]	98 [96,100]	54 [42,64]	72 [66,78]	66 [57,75]	93 [89,97]	76 [69,82]	82 [77,87]
rad. group C	22 [13,32]	100 [100,100]	36 [23,48]	62 [57,67]	66 [56,74]	99 [98,100]	79 [72,85]	84 [80,88]
radiologist average	32 [27,38]	99 [98,100]	48 [42,55]	69 [66,72]	66 [61,71]	95 [93,97]	77 [73,80]	83 [80,86]
ULS	2 [0,6]	100 [100,100]	4 [0,12]	58 [52,64]	12 [6,20]	96 [90,100]	22 [11,33]	54 [46,62]
Merlin (zero-shot)	2 [0,7]	100 [100,100]	4 [0,12]	51 [40,61]	16 [9,23]	96 [90,100]	27 [16,37]	63 [55,72]
R-Super†	49 [38,60]	99 [97,100]	66 [55,75]	90 [85,94]	35 [26,44]	95 [91,98]	50 [39,59]	82 [77,87]

	prostate tumor				adrenal tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	25 [16,34]	98 [96,100]	39 [27,50]	67 [61,72]	0 [0,0]	100 [100,100]	0 [0,0]	54 [52,56]
rad. group B	25 [16,35]	98 [96,100]	39 [27,51]	67 [62,73]	1 [0,2]	100 [100,100]	2 [0,5]	52 [51,54]
rad. group C	10 [4,18]	100 [100,100]	19 [8,30]	57 [54,61]	47 [38,55]	98 [95,100]	63 [54,70]	73 [68,77]
radiologist average	20 [15,25]	99 [98,100]	33 [26,40]	64 [61,67]	16 [12,20]	99 [98,100]	27 [22,32]	60 [58,62]
ULS	12 [3,24]	98 [95,100]	21 [5,38]	60 [49,71]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Merlin (zero-shot)	6 [0,15]	100 [100,100]	11 [0,27]	63 [50,75]	5 [1,9]	100 [100,100]	9 [2,16]	68 [59,76]
R-Super†	54 [43,66]	98 [95,100]	69 [59,78]	90 [85,95]	54 [46,63]	99 [98,100]	70 [62,77]	92 [89,95]

	spleen tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	41 [32,50]	98 [96,100]	58 [48,66]	70 [65,75]	34 [30,37]	98 [98,99]	46 [42,50]	70 [68,72]
rad. group B	41 [32,50]	98 [96,100]	58 [48,66]	70 [65,75]	34 [31,37]	98 [98,99]	47 [43,50]	70 [68,72]
rad. group C	34 [25,42]	100 [100,100]	51 [40,60]	67 [63,72]	28 [25,31]	100 [99,100]	40 [36,44]	65 [64,67]
radiologist average	39 [34,44]	99 [98,100]	56 [50,61]	69 [66,72]	32 [30,34]	99 [98,99]	46 [44,48]	68 [67,69]
ULS	0 [0,0]	98 [95,100]	0 [0,0]	49 [41,58]	7 [4,9]	98 [96,99]	12 [8,15]	54 [51,57]
Merlin (zero-shot)	8 [3,15]	100 [100,100]	16 [5,26]	60 [50,69]	6 [4,9]	99 [98,100]	12 [8,15]	59 [53,65]
R-Super†	67 [58,76]	99 [98,100]	80 [73,86]	86 [80,91]	50 [46,53]	99 [98,99]	64 [60,67]	88 [87,90]

Table D56: **Reader Study (certain level): Unlike R-Super, previous public AI models do not surpass human performance in malignant tumor detection.** In malignant tumor detection (malignant vs no-tumor), we compared the radiologist performance to ULS (segmentation) and Merlin (zero-shot classification), on the reader study test dataset. These are the previous public models that, after R-Super, achieved the highest performance in our experiments (Fig. 2). Unlike R-Super, these previous public AI models were not able to match or surpass the average radiologist sensitivity in any tumor type, at matched specificity. Blue indicates statistical significance ($p < 0.05$), comparing the best previous public model with radiologist average. R-Super (ours, included in the table for reference) was excluded from the comparison. We used DeLong's test for AUC and paired permutation tests (two-sided, 10,000 permutations) for the other metrics. Results are shown as average and 95% confidence intervals (given 1,000 bootstrap samples).

	bladder malignant tumor				esophagus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	46 [30,62]	100 [100,100]	63 [46,77]	78 [69,87]	54 [39,68]	100 [100,100]	70 [56,81]	83 [76,90]
rad. group B	46 [29,62]	100 [100,100]	63 [45,76]	78 [69,86]	54 [38,69]	100 [100,100]	70 [55,82]	83 [75,90]
rad. group C	23 [10,38]	100 [100,100]	37 [19,55]	69 [61,77]	2 [0,8]	100 [100,100]	4 [0,15]	51 [50,54]
radiologist average	38 [29,47]	100 [100,100]	55 [45,64]	75 [70,80]	36 [28,45]	100 [100,100]	53 [44,62]	72 [68,77]
ULS	11 [2,23]	92 [85,98]	18 [4,34]	56 [45,68]	12 [2,22]	100 [100,100]	21 [5,36]	55 [48,63]
Merlin (zero-shot)	9 [0,18]	100 [100,100]	16 [0,30]	59 [47,72]	0 [0,0]	100 [100,100]	0 [0,0]	58 [47,69]
R-Super†	43 [27,59]	100 [100,100]	60 [42,74]	95 [89,99]	67 [53,82]	100 [100,100]	81 [69,90]	96 [93,98]

	gallbladder malignant tumor				uterus malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	60 [40,79]	98 [96,100]	71 [53,86]	82 [72,92]	76 [61,89]	93 [89,97]	75 [61,85]	84 [75,91]
rad. group B	60 [39,80]	98 [96,100]	71 [52,86]	82 [72,92]	76 [61,89]	93 [89,97]	75 [62,84]	83 [75,91]
rad. group C	24 [8,41]	100 [100,100]	39 [14,58]	66 [56,76]	70 [54,85]	99 [98,100]	81 [68,91]	89 [82,96]
radiologist average	48 [37,60]	99 [98,100]	63 [52,72]	77 [71,83]	74 [64,82]	95 [93,97]	76 [68,83]	85 [81,90]
ULS	4 [0,12]	100 [100,100]	8 [0,22]	62 [53,72]	9 [0,20]	96 [90,100]	15 [0,31]	58 [48,69]
Merlin (zero-shot)	4 [0,13]	100 [100,100]	8 [0,23]	50 [37,63]	30 [14,46]	96 [90,100]	44 [23,59]	66 [53,80]
R-Super†	88 [73,100]	99 [97,100]	92 [81,98]	99 [98,100]	64 [47,80]	95 [91,98]	69 [55,81]	86 [78,93]

	prostate malignant tumor				adrenal malignant tumor			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	35 [17,53]	98 [96,100]	49 [27,67]	69 [60,79]	0 [0,0]	100 [100,100]	0 [0,0]	56 [52,62]
rad. group B	35 [18,54]	98 [96,100]	49 [28,67]	69 [61,80]	0 [0,0]	100 [100,100]	0 [0,0]	54 [50,58]
rad. group C	23 [8,40]	100 [100,100]	38 [15,57]	64 [55,73]	36 [21,51]	98 [95,100]	50 [32,65]	67 [59,75]
radiologist average	31 [21,42]	99 [98,100]	45 [33,57]	67 [62,73]	12 [6,19]	99 [98,100]	21 [12,31]	59 [56,63]
ULS	4 [0,13]	98 [95,100]	7 [0,22]	54 [44,66]	0 [0,0]	100 [100,100]	0 [0,0]	50 [50,50]
Merlin (zero-shot)	8 [0,20]	100 [100,100]	14 [0,33]	68 [53,82]	5 [0,13]	100 [100,100]	10 [0,23]	73 [62,83]
R-Super†	65 [46,83]	98 [95,100]	74 [58,86]	92 [83,98]	67 [50,81]	99 [98,100]	79 [66,88]	94 [89,99]

	spleen malignant tumor				**Average**			
	sen.	spe.	F1	AUC	sen.	spe.	F1	AUC
rad. group A	54 [32,77]	98 [96,100]	67 [45,83]	76 [64,88]	46 [40,53]	98 [98,99]	56 [50,61]	76 [72,79]
rad. group B	54 [32,76]	98 [96,100]	67 [45,82]	76 [65,87]	46 [40,52]	98 [98,99]	56 [50,61]	75 [72,78]
rad. group C	41 [21,64]	100 [100,100]	58 [34,78]	70 [60,82]	31 [25,38]	100 [99,100]	44 [36,50]	68 [65,71]
radiologist average	50 [38,61]	99 [98,100]	64 [52,74]	74 [68,81]	41 [37,45]	99 [98,99]	54 [50,58]	73 [71,75]
ULS	0 [0,0]	98 [95,100]	0 [0,0]	58 [46,71]	6 [3,9]	98 [96,99]	10 [5,14]	56 [52,60]
Merlin (zero-shot)	9 [0,24]	100 [100,100]	17 [0,38]	68 [53,81]	9 [6,13]	99 [98,100]	15 [10,21]	63 [56,70]
R-Super†	68 [46,87]	99 [98,100]	79 [61,91]	94 [88,99]	66 [59,72]	99 [98,99]	76 [71,81]	94 [92,96]

Appendix E Training Details

We train using CT patches. For CT-Report pairs, each training patch is designed to capture one target organ. This target organ is randomly chosen, with a high probability of choosing organs with tumors in the report (90%). The training patch fully covers the target organ. Otherwise, a tumor mentioned in the report could fall outside the patch, and the report-based losses would wrongly push the AI to find a tumor that is not visible to the AI.

Training parameters followed the defaults set by MedFormer [28], the segmentation architecture we used inside R-Super. The only new parameters we include are loss weights. We set a loss weight of 1 to the segmentation losses (cross-entropy and Dice, used for CT-Mask pairs), 0.1 for the Volume Loss and Ball Loss, 0.01 for the Attenuation Loss, and 0.1 for the Malignancy Loss. There is no need to carefully tune these weights. For instance, we tuned them once in the UCSF validation dataset, then we used the same weights in all our experiments. We train with AdamW, gradient norm clipping (1), 100 epochs of 1000 batches each, batch size of 2, patch size of 128 x 128 x 128, isotropic voxel spacing of 1 mm, weight decay of 5.00E-2, learning rate of 1.00E-4 (5 epochs of warmup, followed by polynomial decay). CT intensity was clipped between -991 and 500 HU, then normalized. Data augmentation includes rotation, brightness, gamma, contrast, Gaussian blur and Gaussian noise (following MedFormer [28]).

When training with both CT-Report and CT-Mask pairs, we super-sampled the CT-Mask pairs, making them 50% of the samples that the segmentation model saw in each epoch. Augmentation was randomly applied to each sample to reduce the risk of overfitting. When training with CT-Report pairs and only 175 CT-Mask pairs, we super-sampled the CT-Mask pairs to only 10% of the dataset, instead of 50%,. This avoids overfitting with small numbers of masks.

Segmentation models were initialized pre-trained on AbdomenAtlas 3.0 (for the segmentation of organs and tumors in the pancreas, kidney, and liver) [10]. Pre-training followed the same configuration as training (described above), but without CT-Report pairs (nor R-Super losses). Moreover, reports have different levels of precision. Some are coarse, even lacking tumor sizes, while other are very precise, providing the sizes and slices for all tumors. R-Super's losses can exploit the additional information in slices. To maximize their benefit, we first trained R-Super over the whole dataset, including all CT-Report pairs; afterwards, we fine-tuned on only the CT-Report pairs where reports contained slices (same parameters as before, but reducing learning rate to 3.00E-5). Therefore, R-Super training had 3 stages: pre-training on AbdomenAtlas 3.0, training on all CT-Report slices in our dataset, and fine-tuning on the CT-Report pairs with slices. For R-Super†, we included the CT-Mask pairs in these two last training stages. We also included an additional training stage, training with CT-Mask only before training with all CT-Mask and CT-Report pairs. For the task of malignancy identification, R-Super was further fine-tuned on only the cases with pathology reports.

All alternative training methods that we trained (Sec. 2.2) were based on the same architecture as R-Super (MedFormer, or just its encoder in case of the classification baseline). Thus, they used the same training procedure and hyper-parameters as R-Super. Segmentation was pre-trained on AbdomenAtlas 3.0, then fine-tuned on the CT-Mask pairs in our dataset. MTL was pre-trained on AbdomenAtlas 3.0, then fine-tuned on the CT-Mask and CT-Report pairs in our dataset. CLIP was pre-trained (CLIP loss) on the CT-Report pairs in our dataset, then fine-tuned on the CT-Mask pairs. ModelsGenesis was pre-trained on the CT scans in our dataset (masked autoencoder loss), then fine-tuned on the CT-Mask pairs.

As the exception, the nnU-Net [37] was not trained with the same procedure or parameters as the other models, nor the same architecture. The nnU-Net is a self-configuring segmentation framework designed to be used out-of-the-box when employed as a baseline. Thus, we used the standard nnU-Net training procedure, its automatic hyper-parameters, and its ResEncL architecture. We trained it from scratch on the CT-Mask pairs in our dataset. The only non-standard parameter was isotropic 1 mm voxel spacing, as in all other models.

Appendix F VLM Evaluation

This section details how we evaluated the VLMs we compared R-Super to. MedImageInsight was evaluated with zero-shot classification; Lingshu and MedGemma were evaluated with report generation; and Merlin was evaluated with both report generation and zero-shot classification. These evaluation modes were chosen to match the main capabilities and evaluation settings described in each VLM's original code or publication [22–25].

To evaluate report generation, we followed the methodology in [10]: the VLMs generated radiology reports, and we analyzed these reports with an LLM. The LLM determines whether a tumor was detected in each organ. Exactly following [10], we used the Llama 3 70B LLM and the code and prompt from [10]. This LLM and prompt combination were previously validated by radiologists and shown to achieve 96% accuracy for this evaluation task [10].

For zero-shot classification, we computed the cosine similarity between the VLM image embedding of the CT scan and the VLM text embedding of a positive prompt: "organ focal lesion present". We used one positive prompt for each of the 7 organs studied. This cosine similarity was used as the tumor detection score for the organ. Different thresholds over this score define different operating points; Fig. 2 and the tables in Supp. D.6 report

the sensitivity, specificity, F1-Score and balanced accuracy for the operating point with the highest balanced accuracy.

This strategy differs from the standard CLIP zero-shot classification strategy [30]. CLIP typically uses both a positive prompt and a negative prompt, such as "{organ} focal lesion present" and "No {organ} focal lesion". It then computes the similarity between the image embedding and both prompt embeddings, and defines the detection score as the positive-prompt similarity minus the negative-prompt similarity. In our test sets, this standard strategy performed worse than using only the positive-prompt similarity.

Merlin and MedImageInsight required different aggregation strategies because they process CT scans differently. Merlin analyzes a whole 3D CT scan and produces one image embedding. MedImageInsight analyzes the CT scan slice by slice and produces one image embedding per slice. For MedImageInsight, we therefore used the highest cosine similarity between any slice embedding and the positive prompt embedding as the tumor detection score. This strategy performed better than averaging slice scores.

To generate reports with Merlin we used the report generation pipeline released by the model's authors. Merlin received the entire 3D CT scans as input, and it was prompted with "Generate a radiology report for {organ}", with {organ} being one of the 7 organs where we detected tumors. We used one such prompt for each organ, and concatenated the 7 resulting organ-specific reports generated by Merlin. Generation was performed using deterministic decoding (do_sample=False) with beam size 1, a repetition penalty of 1.2, and a maximum generation length of 128 tokens per organ system, consistent with the official Merlin demonstration code.

To generate reports with MedGemma and Lingshu we utilized a slice-based approach, as standard for the models. Each CT case consisted of a set of 2D CT slices (axial) extracted from the original volumetric examination. Following the official inference pipelines for these models, each slice was independently analyzed using a radiology-focused prompt. The prompt requests the models to "*describe the findings visible on the CT scan. Include any abnormality, including tumors if present, for organs that are visible in the scan. Specifically assess and report findings for the following organs when present: esophagus, stomach, duodenum, colon, bladder, liver, pancreas, kidneys, gallbladder, reproductive organs, adrenal glands, and spleen. If an organ is visible and no abnormality is identified, state that it is unremarkable.*" The model generated a report of the visible findings for each slice. The slice-level reports for all slices within a 3D scan were aggregated and provided back to the model in a second-stage summarization step. The model was then instructed to synthesize the slice-level reports into a single structured radiology report representing the entire scan. No image was provided in this second stage, only the slice-level reports.

Appendix G LLM Prompts

The prompt below was used to extract tumor information (size, count, and rough location) from radiology reports, as part of the R-Super training pipeline (Fig. 1). It was created with the assistance of radiologists, and provides information on what should be considered a tumor or not, and how sizes are described in reports. It also includes an example of a report and desired answer (in-context learning).

```
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 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 reports describes a lesion without numeric diameters, but describes the lesion with adjectives
↪  like tiny, small, large, or massive, you **must** include one entry with '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, hypoenchanging, isoenhancing, hererogeneously enhancing, or U (unknown).
↪  The report may use synonyms like hypermetabolic and hypoattenuating, but you must only answer me
↪  hyperenhancing, hypoenchanging, isoenhancing, hererogeneously 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 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.
```

The prompt below is used to extract the tumor z-coordinate (slice) from radiology reports. It also identifies the CT series in which this coordinate was measured, allowing us to choose the correct series to use the slice information in the Ball Loss. Prompt:

```
You are an expert radiologist. Your task is to carefully analyze a radiology report and determine the
↪  image and series numbers for tumors described in the report.

The report may describe one o multiple tumors, and the image and series numbers may appear for all,
↪  none, or only some of the tumors.

To make your task easier, I will give you:

1. **FINDINGS** - the body of the report.
2. **TUMOR LIST** - my pre-extracted tumor descriptions.
   • This list *is* your template: each line already contains
     `series = se` and `image = im`.
   • Replace se and im with the numbers you find in the report.
   • If a number is missing or ambiguous, write u (unknown).
   • Never guess the series or image numbers.

To fill in the TUMOR LIST, you must match each tumor in the TUMOR LIST to tumors described in FINDINGS.
↪  Use all available cues I provide in the list, which may include the tumor organ, location (inside
↪  the organ), size (in mm), and attenuation. Sometimes I may not be able to get some of these cues.
↪  In these cases, I will mark it as 'u' (unknown).
Consider that my list measures tumors in millimetres (mm). Reports use mm if they do not mention the
↪  unit. But they use centimeters (in these cases, they usually write cm). 10 mm = 1 cm.

### Examples of how series / image numbers usually appear in reports"
* `(Se/Im 303/55)` series = 303, image = 55
* `(series 303, image 32)`
* `(3/63)` or `(3:121)` -> series = first number, image = second
* `(image 68)` only -> series = u, image = 68

---

## Worked examples

### Example 1

**REPORT (reduced to only tumor-relevant sentences)**
Liver: Numerous, at least 15, hypoattenuating hepatic lesions nearly all of which are decreased in size
↪  since 12/06. For example, lesion in the dome of segment 7 measures 2.7 x 3.0 cm (Se/Im 4/47),
↪  previously 4.8 x 4.5 cm. Lesion in segment 4A measures 4.4 x 3.4 cm (Se/Im 4/70), previously 5.2 x
↪  4.6 cm.
Spleen: Hypoenhancing lesion in the anterior spleen measures 1.6 x 1.0 cm (Se/Im 4/39) previously 2.0 x
↪  1.1 cm.
Colon: The primary colon cancer, represented by mural thickening of the distal sigmoid colon, is less
↪  conspicuous than 12/06.

**TUMOR LIST TO FILL (input / template)**
tumor id = 1; organ = liver; location = segment 7; size = 27 x 30 mm; attenuation = hypoattenuating;
↪  series = se; image = im;
tumor id = 2; organ = liver; location = segment 4; size = 44 x 34 mm; attenuation = hypoattenuating;
↪  series = se; image = im;
tumor id = 3; organ = spleen; location = u; size = 16 x 10 mm; attenuation = hypoattenuating; series =
↪  se; image = im;
tumor id = 4; organ = colon; location = u; size = u; attenuation = u; series = se; image = im;

**FILLED LIST (expected model output)**
tumor id = 1; organ = liver; location = segment 7; size = 27 x 30 mm; attenuation = hypoattenuating;
↪  series = 4; image = 47;
```

```
tumor id = 2; organ = liver; location = segment 4; size = 44 x 34 mm; attenuation = hypoattenuating;
↪  series = 4; image = 70;
tumor id = 3; organ = spleen; location = u; size = 16 x 10 mm; attenuation = hypoattenuating; series =
↪  4; image = 39;
tumor id = 4; organ = colon; location = u; size = u; attenuation = u; series = u; image = u;

**JUSTIFICATION**
tumor id = 1; series = 4, image = 47 - extracted from: "lesion in the dome of segment 7 measures 2.7 x
↪  3.0 cm (Se/Im 4/47)"; matched by organ (liver), location (segment 7), and size = 27 x 30 mm.
tumor id = 2; series = 4, image = 70 - extracted from: "lesion in segment 4A measures 4.4 x 3.4 cm
↪  (Se/Im 4/70)"; matched by organ (liver), location (segment 4), and size = 44 x 34 mm.
tumor id = 3; series = 4, image = 39 - extracted from: "hypoenhancing lesion in the anterior spleen
↪  measures 1.6 x 1.0 cm (Se/Im 4/39)"; matched by organ (spleen) and size = 16 x 10 mm.
tumor id = 4; series = u, image = u - no series / image numbers are given for the sigmoid-colon tumour;
↪  therefore both values are unknown.

---

### Example 2

**REPORT (reduced to only tumor-relevant sentences)**
Spleen: Slightly decreased size of a hypodense lesion in the spleen measuring up to 3.1 cm (series 5,
↪  image 30) previously 2.7 cm. Unchanged small hypervascular nodule at the tip of the spleen (series
↪  601, image 43).
Adrenal Glands: Stable 1.5 cm left adrenal nodule (series 5, image 40).
Kidneys: Unchanged 1.3 cm enhancing lesion in the inferior pole of the right kidney (series 5, image
↪  79). Additional bilateral sub-centimeter renal hypodensities that are too small to characterize,
↪  stable from prior.
Peritoneum: Mild omental fat stranding along the left anterior abdomen (series 5, image 100), new from
↪  prior. Similar sequela of prior epiploic appendagitis anterior to the descending colon (series 5,
↪  image 92). Similar small capsular nodule adjacent to the inferior spleen (series 5, image 55).
Reproductive organs: Slightly increased size of a heterogeneously enhancing mass in the uterus, now
↪  measuring 6.5 x 5.1 cm (series 5, image 133), previously 6.1 x 4.8 cm.
Liver: Unchanged sub-centimeter hypodensities that are too small to characterize.

**TUMOUR LIST TO FILL (input / template)**
tumour id = 1; organ = spleen; location = u; size = 31 mm; attenuation = hypoattenuating; series = se;
↪  image = im;
tumour id = 2; organ = spleen; location = u; size = u; attenuation = hyperattenuating; series = se;
↪  image = im;
tumour id = 3; organ = adrenal gland; location = left; size = 15 mm; attenuation = u; series = se;
↪  image = im;
tumour id = 4; organ = kidney; location = right; size = 13 mm; attenuation = hyperattenuating; series =
↪  se; image = im;
tumour id = 5; organ = uterus; location = u; size = 65 x 51 mm; attenuation = u; series = se; image =
↪  im;

**FILLED LIST (expected model output)**
tumour id = 1; organ = spleen; location = u; size = 31 mm; attenuation = hypoattenuating; series = 5;
↪  image = 30;
tumour id = 2; organ = spleen; location = u; size = u; attenuation = hyperattenuating; series = 601;
↪  image = 43;
tumour id = 3; organ = adrenal gland; location = left; size = 15 mm; attenuation = u; series = 5; image
↪  = 40;
tumour id = 4; organ = kidney; location = right; size = 13 mm; attenuation = hyperattenuating; series =
↪  5; image = 79;
tumour id = 5; organ = uterus; location = u; size = 65 x 51 mm; attenuation = ug; series = 5; image =
↪  133;

**JUSTIFICATION**
tumor id = 1; series = 5, image = 30 - extracted from: "hypodense lesion in the spleen measuring up to
↪  3.1 cm (series 5, image 30)"; matched by organ (spleen) and size = 31 mm.
tumor id = 2; series = 601, image = 43 - extracted from: "hypervascular nodule at the tip of the spleen
↪  (series 601, image 43)"; matched by organ (spleen) and hyperattenuating (hypervascular)
↪  attenuation.
tumor id = 3; series = 5, image = 40 - extracted from: "stable 1.5 cm left adrenal nodule (series 5,
↪  image 40)"; matched by organ (adrenal), location (left), and size = 15 mm.
tumor id = 4; series = 5, image = 79 - extracted from: "1.3 cm enhancing lesion in the inferior pole of
↪  the right kidney (series 5, image 79)"; matched by organ (kidney), location (right), enhancing
↪  nature, and size = 13 mm.
tumor id = 5; series = 5, image = 133 - extracted from: "heterogeneously enhancing mass in the uterus
↪  ... 6.5 x 5.1 cm (series 5, image 133)"; matched by organ (uterus) and size = 65 x 51 mm.



---

### Example 3

**REPORT (reduced to only tumor-relevant sentences)**
```

```
Liver:  Diffusely hypoattenuating. Unchanged subcentimeter segment 2 hyperdense focus (3/32), likely a
↪  benign hemangioma.
Gallbladder: Cholelithiasis without evidence of acute cholecystitis.
Spleen:  Similar-appearing hypoattenuating lesions, some of which are too small to characterize, which
↪  have been gradually increasing in size since at least 9/13/2021. The largest measures up to 1.8 cm.
Pancreas:  Unremarkable

**TUMOUR LIST TO FILL (input / template)**
tumor id = tumor 3; organ = spleen; location = u; size = 18.0 mm; attenuation = hypoattenuating; series
↪  = se; image = im;

**FILLED LIST (expected model output)**
tumor id = tumor 3; organ = spleen; location = u; size = 18.0 mm; attenuation = hypoattenuating; series
↪  = u; image = u;

**JUSTIFICATION**
tumor id = tumor 3; series = u, image = u - The rules ask me to check the tumor organ and size when
↪  matching tumors in the TUMOUR LIST TO FILL and in FINDINGS. tumor 3 in the list is a spleen tumor
↪  measuring 18mm. The FINDINGS mentions a spleen tumor with 18 mm, but does not have series number or
↪  image number for it: 'Spleen:  Similar-appearing hypoattenuating lesions, some of which are too
↪  small to characterize, which have been gradually increasing in size since at least 9/13/2021. The
↪  largest measures up to 1.8 cm.'. Therefore, I wrote series = u; image = u for tumor 3.

---

## YOUR ACTUAL TASK

Here is the FINDINGS section of the report to analyze:
%(findings_section)s

Here is the TUMOR LIST to fill (replace ‘‘se’’ and ‘‘im’’):
%(tumors_list)s

After filling the list, provide a short justification for each tumor. The justification must explain
↪  the matching you did, and where in the findings you found the tumor organ and tumor size for each
↪  tumor in my list.
Quote the sentence(s) where you found the numbers, or explain why they are unknown.
```

We use pathology reports to determine whether tumors are malignant or benign. To this end, we use the prompt below. It was also developed with the assistance of radiologists, and provides guidelines to the LLM on how do determine malignancy.

```
I am sending you below many pathology reports, all for the same patient.

YOUR TASK
Read ALL reports carefully and produce SIX lists:
1. organs_with_primary_malignant_tumors
2. organs_with_metastatic_tumors
3. organs_with_benign_tumors
4. organs_with_unknown_tumor_type
5. malignant_tumor_types
6. benign_tumor_types

The ‘‘tumor types’’ lists should specify, for each organ, the type of tumor identified (e.g.,
↪  adenocarcinoma, squamous cell carcinoma, renal cell carcinoma, lipoma, hemangioma, cyst, abscess).
↪  If an organ has multiple types, list all separated by semicolons.

Then provide a justification for each organ listed, citing exact evidence phrases (quoted).

---

### OUTPUT FORMAT (strict; no extra prose before/after)

organs_with_primary_malignant_tumors: [<comma-separated organs>]
organs_with_metastatic_tumors: [<comma-separated organs>]
organs_with_benign_tumors: [<comma-separated organs>]
organs_with_unknown_tumor_type: [<comma-separated organs>]
malignant_tumor_types: {<organ>: <tumor type(s)>, ...}
benign_tumor_types: {<organ>: <tumor type(s)>, ...}

Justification:
- <organ>: <label> - Evidence: "<verbatim phrase>". Rationale: <1-4 lines>.
- ...

---

### DEFINITIONS (be conservative)

**Primary malignant tumor** = malignant neoplasm originating in that organ.
```

```
If a report or history states a prior primary (even if ‘‘no residual tumor’’), include that organ as
↪  primary and note ‘‘history only’’.

**Metastatic tumor** = malignant tumor found in an organ but arising elsewhere (e.g., ‘‘metastatic
↪  carcinoma consistent with lung primary’’).

**Benign tumor/lesion** = explicitly benign lesion (e.g., ‘‘benign cyst’’, ‘‘lipoma’’, ‘‘adenoma’’,
↪  ‘‘hemangioma’’, ‘‘granuloma’’, ‘‘abscess’’, ‘‘fibroadenoma’’, ‘‘leiomyoma’’) OR a named
↪  mass/nodule/cyst/polyp described as benign.

---

### BENIGN INCLUSION RULE (must meet BOTH)
A) Mentions a **lesion noun**: {‘‘lesion’’, ‘‘mass’’, ‘‘nodule’’, ‘‘tumor’’, ‘‘neoplasm’’, ‘‘cyst’’, ‘‘polyp’’,
↪  ‘‘adenoma’’, ‘‘lipoma’’, ‘‘hemangioma’’, ‘‘leiomyoma’’, ‘‘hamartoma’’, ‘‘abscess’’, ‘‘granuloma’’}.
B) The lesion is described as benign or is canonically benign.

### BENIGN EXCLUSION RULE
Do NOT include normal or uninvolved tissues:
- ‘‘Benign <normal tissue>’’ (e.g., ‘‘benign skeletal muscle’’, ‘‘benign submandibular gland’’) → exclude.
- ‘‘No tumor identified’’, ‘‘negative for malignancy’’, ‘‘benign mucosa/fat/muscle/lymph node’’ → exclude.
- Reactive or inflammatory processes (e.g., esophagitis, fibrosis, metaplasia, cholecystitis) without a
↪  mass → exclude.
- Fungal or bacterial infection without a lesion noun → exclude.

---

### UNKNOWN BUCKET
Use UNKNOWN for:
- ‘‘Atypical’’, ‘‘suspicious’’, ‘‘cannot rule out’’, ‘‘insufficient for diagnosis’’, or contradictory results
↪  across reports.

---

### CANONICAL ORGAN NAMES
Use exactly these names when applicable:
uterus, prostate, right kidney, left kidney, right adrenal gland, left adrenal gland, pancreas, liver,
↪  spleen, bladder, gallbladder, duodenum, stomach, colon, esophagus.
If the tumor is inside one of these, use the canonical name (e.g., use *uterus* for endometrial
↪  carcinoma).
For others (e.g., lung, skin, lymph node), use the report’s term.

---

### EVIDENCE & JUSTIFICATION
- For EACH organ listed, provide one quoted phrase that proves the classification.
- If classification is based on history, quote that phrase and mark ‘‘history only’’.
- Include specimen/part/date if available.
- If the same organ has malignant and benign lesions, include it in both lists and explain.

---

### QUALITY CHECKS
- Prefer UNKNOWN over guessing.
- Do NOT treat normal tissue or ‘‘benign <tissue>’’ as a benign tumor.
- Be explicit about tumor types; if a lesion type is uncertain, label it as ‘‘unknown type’’ and explain.
- Justify every entry with a phrase containing a lesion noun and diagnostic qualifier
↪  (benign/malignant/metastatic).

---

Now read the reports below and output ONLY the specified blocks.
```

Appendix H Instructions for Reader Study

The instructions below were provided to the radiologists before the reader study. Radiologists did not know the AI model outputs, nor the diagnoses of the patients.

```
Reader Study Instructions

Analyze the entire CT scan. For each CT, answer the following questions:

Which organs certainly have a lesion (benign or malignant tumor)?

Write the names of the organs that for sure have a focal lesion.

Definition of focal lesion: discrete mass-like focus, cyst, solid lesion, enhancing nodule, focal wall
↪ thickening, and focal (not diffuse) hyper/hypo densities that you would mention if you were writing
↪ a report for this CT. Include any lesion that may be a primary cancer, metastasis, benign lesion,
↪ or cyst.

You can ignore: bone lesions, diffuse fatty change, diverticulum (unless it is suspicious for
↪ malignancy), stones (e.g., renal/gallbladder stones), perfusion anomalies unless mass-like,
↪ non-focal inflammation, postoperative change unless mass-like, and simple fluid collections unless
↪ clearly cystic lesion of an organ (e.g., abscess, or postoperative changes).

Which organs possibly have a lesion (benign or malignant tumor)?

Here, write the names of the organs that possibly have a focal lesion, but you are not sure if there is
↪ a lesion there or no lesion. Consider the same focal lesion definition as in (1).

Which organs certainly have a malignant lesion (primary or metastasis)?

Here, write the names of the organs that for sure have a malignant lesion. This lesion can be a primary
↪ cancer or metastasis. Not a benign lesion or cyst.

Which organs possibly have a malignant lesion (primary or metastasis)?

Here, write the names of the organs that possibly have a malignant lesion. Consider any lesion that is
↪ suspicious, which makes you concerned for a possible malignancy.

How to prepare your answer?
We provide a CSV file, called answer_sheet.xlsx. It has the CT scan IDs and the 4 questions above (4
↪ columns). You can answer each question with just organ names. When a CT has lesions in multiple
↪ organs, you can separate the organ names with commas. When a CT has no lesions, leave its row blank
↪ (no need to write anything).

Analyze the whole scan, and be sure to review the following organs (if present): spleen, bladder,
↪ gallbladder, uterus, prostate, esophagus, stomach, duodenum, intestines, liver, pancreas, kidneys,
↪ adrenal glands.
```

CT	Which organs certainly have a lesion (benign or malignant tumor) for sure?	Which organs possibly have a lesion (benign or malignant tumor)?
XX	pancreas	pancreas, spleen

Which organs certainly have a malignant lesion (primary or metastasis)?	Which organs possibly have a malignant lesion (primary or metastasis)?
pancreas	pancreas

```
This example indicates that lesions exist for sure in the spleen, bladder and stomach. The spleen
↪ lesion is for sure malignant. The bladder lesion is possibly malignant (unsure). The stomach lesion
↪ is for sure benign.

Loading data:
Our data is provided in the nifti format. It can be opened with softwares like ITK-SNAP.
ITK-SNAP can be downloaded here:
https://www.itksnap.org/pmwiki/pmwiki.php?n=Downloads.SNAP4
After opening ITK-SNAP, you can load an image by clicking File -> Open Main Image ...
You can adjust the CT contrast window by clicking here:
```

Then clicking on contrast.
In the example below, we adjusted contrast between -200 and 200 HU.