

CAPTOR: a vision-language foundation model for Chest-Abdomen-Pelvis Oncologic CT

S  l  na Jean-Mactoux^{1,2}Sophie Loizillon²Ghina Jardali^{1,3}Amandine Cromb  ^{3,4,5}Nathalie Lassau^{1,3}Alexandre B  ne²¹ Biomaps, UMR1281, INSERM, CNRS, CEA, Universit   Paris-Saclay, Villejuif, France² Guerbet Research, Villepinte, France³ Department of Diagnostic Oncologic Imaging, Gustave Roussy Institute, Villejuif, France⁴ Department of Radiology, Pellegrin University Hospital, Bordeaux, France⁵ Bordeaux Institute of Oncology, BRIC U1312, INSERM, Team SARCOTARGET, Bordeaux, France

Abstract

Purpose

To develop and evaluate CAPTOR, a deep-learning vision-language foundation model for oncologic chest-abdomen-pelvis (CAP) CT interpretation.

Materials & Methods

CAPTOR (Chest-Abdomen-Pelvis foundation model for Oncologic Radiology) is a contrastive vision-language model, trained using a pretraining strategy for oncologic CAP CT (in single- or multiple-series acquisitions), on a multicenter retrospective dataset from North America, Europe and Asia, predominantly from tumor lesion examinations. CAPTOR was compared with state-of-the-art CT foundation models on three downstream tasks: disease detection, survival prediction using TumorImagingBench (1) and image retrieval. Classification and retrieval were both evaluated on held-out internal and public test sets, stratified by anatomical coverage and indication. Bootstrap resampling was used to estimate confidence intervals, and pairwise comparisons were performed using permutation tests.

Results

CAPTOR was trained on 25,512 CT-report pairs (median age 66 years, 49% female) and evaluated on 10,794 studies (median age 60 years, 52% female). TumorImagingBench datasets comprised between 133 and 421 studies each (median ages between 58.5 and 69 years, between 25% and 41% female each). CAPTOR outperformed all baseline models, on the classification task, on the CAP subset (AUC 0.767, 95%CI: 0.734-0.793, vs. 0.695, 95%CI: 0.645-0.736 for the

next-best, $p < 0.001$) and oncologic subset (AUC 0.784, 95%CI: 0.759-0.805, vs. 0.734, 95%CI: 0.703-0.767 $p < 0.001$), ranked first on three of the four oncologic prognostic benchmarks, and achieved superior retrieval performance on CAP and oncologic subsets.

Conclusion

A foundation model trained on oncology-enriched CAP CT data outperformed relevant state-of-the-art models in its target clinical context, underscoring the importance of clinically aligned pretraining.

We release our preprocessing, training and inference code at this url : <https://github.com/Guerbet-AI/CAPTOR> and the model weights at this url : <https://huggingface.co/Guerbet-AI/CAPTOR>

1 Introduction

With a projected increase of 63% in cancer incidence by 2045 [1], pressure intensifies on radiological practice, especially for Chest-Abdomen-Pelvis (CAP) CT interpretation, a key modality for oncologic staging, metastatic assessment, and treatment monitoring [2, 3, 4, 5]. This further strains the radiology workforce, with a radiologist shortage documented across Europe and the United States [6, 7], and elevates the risk of interpretative errors, particularly under fatigue like during overnight shifts [8]. Scalable AI tools supporting multiple downstream tasks from a single examination are increasingly recognized by major radiology societies for their workflow augmentation potential [9].

Recent advances on foundation models, and specif-

ically vision-language models (VLMs) enable joint image-text representation learning from CT images and radiology reports through contrastive pretraining, without manual annotations. CLIP [10] methods applied to medical context, such as CT-CLIP [11] or Merlin [12] have demonstrated the potential of this approach across various diagnostic and prognostic tasks.

However, most CT foundation models are trained on anatomically specific datasets (e.g. Chest for CT-CLIP or Abdomen-Pelvis for Merlin), limiting their applicability to full CAP protocols. Others, such as SPECTRE [13] and COLIPRI [14], both transformer-based multimodal foundation models, achieve broader anatomical coverage through heterogeneous training sets, but without oncologic-rich training data. While oncologic-specific foundation models exist, they primarily target lesion detection, characterization, or treatment response rather than whole-image analysis [15, 16]. Yet, radiologists interpreting CAP CT scans in oncologic patients must also identify a wide spectrum of non-cancerous findings, ranging from treatment-related complications to age- and comorbidity-related conditions, as well as incidental findings requiring further evaluation, which can be overlooked under high workload and when the clinical question is focused on cancer. A gap therefore exists between narrowly focused oncologic models, and generalist foundation models lacking exposure to multimorbidity-rich oncologic CAP imaging, with no evidence that oncology patients require dedicated foundation models.

To bridge this gap, we developed CAPTOR (Chest-Abdomen-Pelvis foundaTion model for Oncologic Radiology), a contrastive vision-language foundation model trained using a preprocessing strategy to handle whole CAP CT examinations, in single or multiple-series acquisitions. The training set comprised 25,512 predominantly oncologic CAP CT studies paired with radiology reports, from diverse scanners and protocols, reflecting routine clinical practice. CAPTOR was compared with five state-of-the-art CT foundation models, including region-specific (CT-CLIP [11], Merlin [12]) and generalist models (SPECTRE [13], COLIPRI [14], CT-FM [17]) across three downstream tasks : medical condition classification, 2-year survival prediction using TumorImagingBench [18], and image retrieval. To assess the influence of pretraining dataset size, we additionally evaluated a larger variant trained on approximately

100,000 studies (CAPTOR-100K).

2 Materials and Methods

2.1 Datasets

This retrospective study comprised three dataset categories: (i) an internal multicenter image-report dataset for vision-language pretraining, collected under data use agreements with 11 institutions across the United States (66%), Spain (16%), France (14%), Japan (3%) and Vietnam (1%), (ii) public image-report datasets for downstream experiments : Merlin [12], CT-RATE [11], INSPECT [19] and BIMCV-R [20], and (iii) public oncologic benchmark datasets for two-year post-treatment survival prediction, using TumorImagingBench [18]: NSCLC-Radiomics [21] and NSCLC-Radiogenomics [22] on chest CT, and C4KC-KiTS [23] and Colorectal-Liver-Metastases [24] on abdominal CT. All datasets were used under their respective use terms. All data were de-identified; informed consent was waived by the local institutional review board or ethics committee and data handling complied with applicable regional data protection regulations, including HIPAA and GDPR where relevant.

The internal dataset was curated specifically for oncologic CT interpretation, predominantly from tumor lesion examinations, plus data for incidental pancreatic cancer detection. Studies' indication (oncologic vs. non-oncologic) was inferred from radiology reports using the LLM Mistral Large 3 [25]. A random patient-level 80%/20% split was used to create training and testing sets.

Appendices A.1-A.3 provide details on all datasets

2.2 CAPTOR Foundation Model – Architecture

We used the Merlin [12] architecture, a contrastive VLM using Clinical Longformer [26] as text encoder, and an inflated ResNet152 [27] initialized with 2D ImageNet pretrained weights as the 3D image encoder.

2.3 CAPTOR Foundation Model – Preprocessing

Figure 1 provides an overview of the CAPTOR pipeline, encompassing data preprocessing, model training and evaluation.

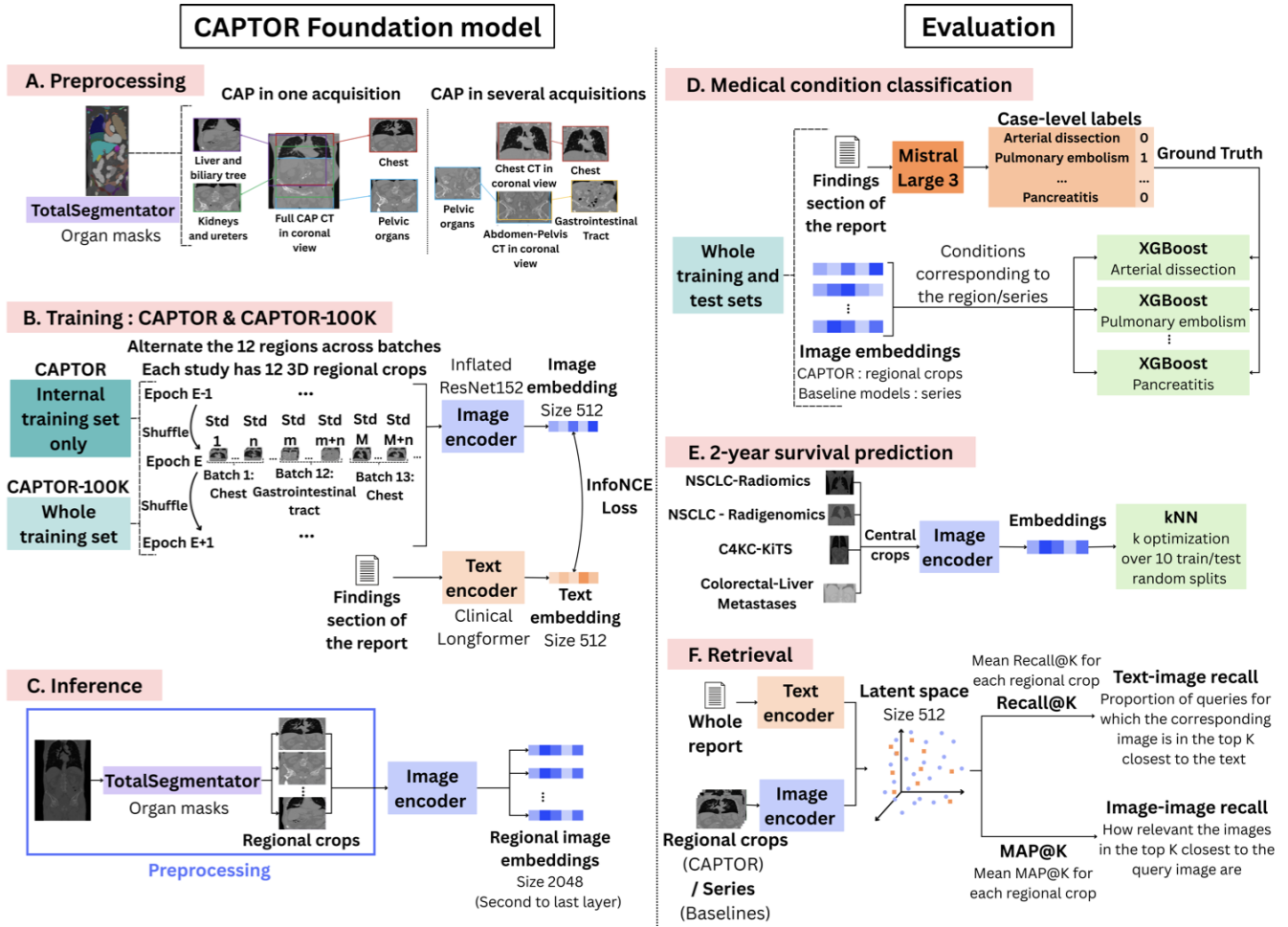


Figure 1: Overview of the CAPTOR pipeline. CT volumes are cropped into 12 anatomical sub-regions using TotalSegmentator28 masks (A); regional crops are paired with the Findings section of the corresponding report and CAPTOR is trained with a rotating-region contrastive objective (B). At inference, TotalSegmentator masks are used to generate regional crops, and each CT yields one embedding per region (C). The resulting frozen image and text encoders are evaluated via XGBoost classification, k-nearest neighbor (kNN) survival prediction, and cross-modal retrieval (D, E, F). MAP@K = mean average precision at K, with K the number of retrieved images, std = Study

CAPTOR’s main methodological innovation lies in its preprocessing strategy (Fig. 1.A) : rather than using fixed central crops like Merlin [12], we use TotalSegmentator [28] masks to center each crop on a specific organ or anatomical region, enabling a region-specific architecture to handle whole CAP CT examinations, including multiple-series studies (Appendix A.4).

2.4 CAPTOR Foundation Model – Training

CAPTOR was trained on the proprietary dataset with the contrastive InfoNCE loss, pairing region-centered crops with the *Findings* part of the reports.

Batches comprise studies, with the anatomical region rotating each batch (Fig. 1B). Shuffling between epochs ensures all crops of each study are seen during training. This strategy allows the model to learn representations from complete, including multi-series, CAP examinations, without modifying the underlying architecture (hyperparameters in Appendix A.5).

2.5 Evaluation – Medical condition classification

We selected 26 medical conditions relevant to oncologic CAP CT interpretation from an initial list of 178, by severity and utility as detection targets.

Reference labels were built from radiology reports

using Mistral Large 3 [25], a method shown to achieve high agreement with expert annotations and provide useful supervision at scale [29, 30]. To verify extraction quality, we compared performance against GPT-5.4 [31] and GPT-4o mini [32] on CT-RATE’s test set using the original labels and we additionally manually reviewed 25 randomly selected reports.

For each condition, an XGBoost classifier was trained on embeddings from the relevant anatomical crop (Fig. 1.D). Performance was assessed on subsets stratified by anatomical coverage (Chest, AP or CAP) and indication (oncological vs. non-oncological), using the macro-AUC (unweighted mean of per-condition area under the receiver operating characteristic curve), restricted to conditions relevant to each subset.

Generalizability to non-oncologic data was evaluated on CT-RATE [11] and RAD-ChestCT [33], both chest CT datasets, using the original labels for condition classification.

Additional details in Appendices A.7-A.9.

2.6 Evaluation – Survival prediction benchmark

We also evaluated all models on four two-year survival prediction tasks from TumorImagingBench [18], using each model’s own preprocessing. Following TumorImagingBench’s method, kNN classifiers (k optimized in {1-50}) were fit per model and dataset over 10 random splits (Fig. 1.E).

2.7 Evaluation – Retrieval

Retrieval tasks evaluate embedding-space quality by measuring whether semantically similar images and reports are mapped close together. Zero-shot text-image retrieval was evaluated using Recall@K, the proportion of text queries whose matching CT examination was retrieved among the top K images, with K the number of retrieved items considered. Zero-shot image-image retrieval was evaluated using Mean Average Precision at K (MAP@K) [34], rewarding early retrieval of images sharing at least one positive condition label (Fig. 1.F and full metrics definitions in Appendix A.10). Fold-increase over the random baseline was used for both metrics, better accounting for varying subset and K difficulty than raw scores. SPECTRE [13] could not be evaluated (unavailable projection-layer weights), and CT-FM [17], a vision-only model, was excluded from text-image retrieval.

2.8 Evaluation - Impact of training data volume

To assess the effect of larger, more heterogeneous, less clinically aligned training data, CAPTOR-100K was trained on the complete training set (internal and public combined) and compared with CAPTOR (internal only) and baseline models. To study training data volume’s impact on classification performance, we trained the classifiers with nested subsets of the whole training set (10%, 25%, 50% and 75%), generated using iterative multilabel stratified sampling.

2.9 Statistical analysis

We compared CAPTOR with five state-of-the-art foundation models : Merlin [12], CT-CLIP [11], SPECTRE [13], COLIPRI [14] and CT-FM [17] (Appendix A.6, Table 4). For all tasks, 95% confidence intervals were estimated with a patient-level bootstrap test (2,000 resamples; 1,000 for retrieval). For classification, per-condition pairwise comparisons were performed using the DeLong test. Otherwise, pairwise comparison significance was assessed with a patient-level two-sided permutation test (2,000 permutations). All p-values were corrected for each task separately, using the Benjamini-Hochberg false discovery rate (FDR) procedure, and differences were considered significant with a corrected p-value below 0.05.

All analyses were performed using Python version 3.12.12 (Python Software Foundation); Torch version 2.7.0; NumPy version 1.26.4; Pandas version 2.2.3, and scikit-learn version 1.8.0.

2.10 Code availability

We release our preprocessing, training and inference code at this url : <https://github.com/Guerbet-AI/CAPTOR> and the model weights at this url : <https://huggingface.co/Guerbet-AI/CAPTOR>

3 Results

3.1 Dataset characteristics

Table 1 summarizes dataset demographics and Figure 2 outlines how each dataset is used for pretraining and downstream tasks. The internal dataset comprised 29,893 studies (32,689 series) from 11 providers across the United States (66%), Spain (16%), France (14%),

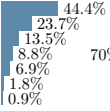
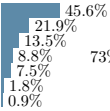
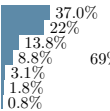
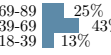
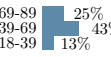
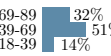
	Dataset	Subset	Studies	Age	Sex	Body Region	Indication	Contrast	Provenance	Report Section(s)
Private	Internal dataset	Total	29,893 studies 32,689 series	66.0 [56.0-75.0]	F : 49% M : 51%		70% oncologic	Injected : 70% Non-injected : 16% Unknown : 14%	Multi-centric from 11 providers in : United States (66%) Spain (16%) France (14%) Japan (3%) Vietnam (1%)	Full report
		Train	25,512 studies 27,698 series	66.0 [56.0-75.0]	F : 49% M : 51%		73% oncologic	Injected : 67% Non-injected : 19% Unknown : 13%		
		Test	4,381 studies 4,991 series	68.0 [57.0-77.0]	F : 48% M : 52%		69% oncologic	Injected : 66% Non-injected : 17% Unknown : 17%		
Public	CT-RATE	Total	21,304 studies 25,692 series 50,188 volumes	46.0 [35.0-62.0]	F : 42% M : 58%	Chest	Unknown	Non-injected	Mono-centric from Istanbul Medipol University Mega Hospital	Findings & Impressions
		Train	20,000 studies 24,128 series 47,149 volumes	45.0 [34.0-61.0]	F : 42% M : 58%					
		Test	1,304 studies 1,304 series 3,039 volumes	45.0 [34.0-60.0]	F : 44% M : 56%					
	Merlin	Total	25,414 studies	55.0 [39.0-68.0]	F : 56% M : 44%	Abdomen & Pelvis	Unknown	Injected	Multi-centric from Stanford Research Repository	Findings & Impressions
		Train	20,305 studies	55.0 [39.0-68.0]	F : 56% M : 44%					
		Test	5,109 studies	55.0 [40.0-68.0]	F : 57% M : 43%					
	INSPECT (CTPA)	Total	23,248 studies		F : 55% M : 45%	Chest	Pulmonary Embolism	Injected	Multi-centric from Stanford Research Repository	Impressions only
		Train	20,034 studies		F : 54% M : 46%					
		Test	3,214 studies		F : 56% M : 44%					
	BIMCV-R	Total	8,069	Unknown	Unknown	Chest	Unknown	Unknown	Multi-centric from Valencian Region Medical ImageBank	Full report
Public TumorImagingBench	NSCLC - Radiomics	Total	421 studies	68.7 [61.2-75.9]	F : 31% M : 69%	Chest	NSCLC patients : - pre-radiation CT - GTV expert segmentation	Unknown	MAASTRO Clinic, The Netherlands	No reports
	NSCLC - Radiogenomics	Total	133 studies	69.0 [64.0-76.0]	F : 25% M : 75%	Chest	NSCLC patients : - pre-operative CT - GTV expert segmentation	Unknown	Stanford University School of Medicine and Palo Alto Veterans Affairs Healthcare System	No reports
	C4KC-KiTS	Total	134 studies	58.5 [50.0-68.0]	F : 41% M : 59%	Abdomen	Renal tumor patients : - partial/radical nephrectomy - pre-operative CT - tumor/kidney segmentations	Injected	University of Minnesota Medical Center	No reports
	Colorectal-Liver Metastases	Total	194 studies	61.0 [52.0-68.8]	F : 40% M : 60%	Abdomen	CRLM patients : -pre-operative CT - liver/tumor /vessels segmentation - CRLM resection	Injected	Memorial Sloan Kettering Cancer Center	No reports

Table 1: Composition of the internal and public datasets used for training and testing. Percentages are reported at the study level. Train/test sets for TumorImagingBench1 are not included, as their original training/test splits were not used: survival prediction was instead assessed via k-nearest-neighbor classifiers across 10 random splits (Materials and Methods). The median age is shown, with inter-quartile range in brackets. CTPA = CT pulmonary angiogram; F=Female, M=Male, NSCLC = non-small cell lung cancer; GTV = gross tumor volume; CRLM = colorectal liver metastases.

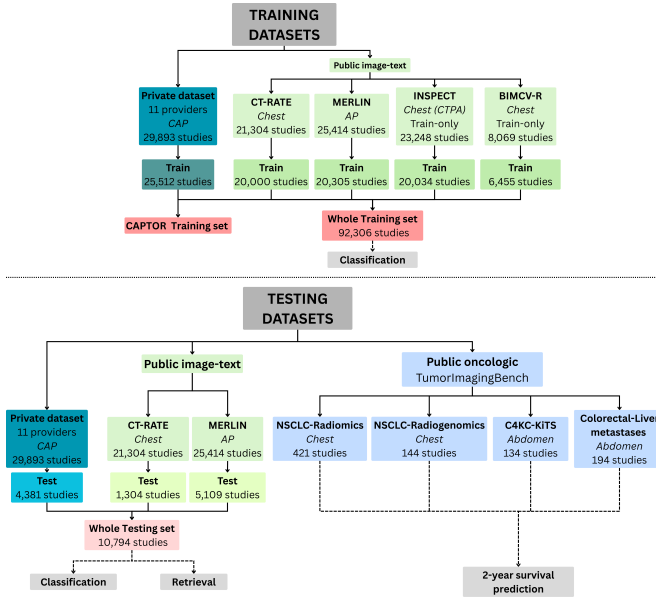


Figure 2: Flow chart showing the datasets used for training CAPTOR and for training/testing downstream tasks. CAP= Chest-Abdomen-Pelvis, AP= Abdomen-Pelvis

Japan (3%) and Vietnam (1%), with a median age of 66.0 years, 49% female. Detailed characteristics by country of the proprietary dataset in Appendix B.1. The public image-text datasets included CT-RATE [11] (21,304 chest CT studies), Merlin [12] (25,414 abdomen-pelvis CT studies), INSPECT [19] (23,248 CT pulmonary angiograms), and BIMCV-R [20] (8,069 chest CT studies). The resulting full training and testing sets including internal and public datasets contain 92,302 studies and 10,794 studies respectively. The testing set has a median age of 60 years [IQR 45-73] and was 52% female.

The four TumorImagingBench [18] datasets comprised between 134 and 421 studies each, covering lung (NSCLC-Radiomics [21]: 421 studies, NSCLC-Radiogenomics [22]: 133 studies) and abdominal oncologic cohorts (C4KC-KiTS [23]: 134 studies; Colorectal-Liver Metastases [24]: 194 studies), all without associated radiology reports.

3.2 Medical condition classification

Automatic performance assessment of label extraction method on CT-RATE [11], showed that GPT-5.4 [31], Mistral Large 3 [25], and GPT-4o mini [32] achieved F1-scores of 0.917, 0.915, and 0.891 respectively (Appendix B.2, Figure 7). Mistral Large 3 was selected for its comparable performance for approxi-

mately 13% of the inference cost.

Manual review of 25 randomly selected reports identified only two extraction errors; both times it was the coronary artery disease label set to True because the report mentioned atherosclerosis.

As shown in Figure 3, CAPTOR achieved the highest overall performance, reaching an AUC of 0.849 (95%CI: 0.828-0.866) on the whole test set, significantly ($p < 0.001$) outperforming all models except for Merlin [12] (0.845, 95% CI: 0.817 – 0.859).

Across anatomical regions, CAPTOR improved over CT-CLIP [11] on the Chest subset (0.838, 95%CI: 0.789-0.865 vs. 0.732, 95%CI: 0.674 – 0.777, $p < 0.001$), while Merlin [12] outperformed CAPTOR on the AP subset (0.807, 95%CI: 0.797 – 0.816 vs. 0.821, 95%CI: 0.812-0.829, $p < 0.01$). On the non-oncological subset, CAPTOR was only statistically significantly outperformed by SPECTRE [13] (0.859, 95%CI: 0.833 – 0.879 vs. 0.874, 95%CI: 0.848 – 0.892, $p < 0.01$) despite its oncology-focused training.

Focusing specifically on our target use case, CAP images in an oncological context, CAPTOR statistically significantly outperformed all baseline models, reaching an AUC of 0.767 (95%CI: 0.734 – 0.793) on the CAP subset and 0.784 (95%CI: 0.759 – 0.805) on the oncological subset. Notably, Merlin outperformed the generalist models on both subsets despite being trained only on AP images. Figure 4 shows the details of per-condition classification performance on the oncological subset for CAPTOR and all baselines. Per-condition proportions and significance counts are available in Appendix B.3-B.4.

CAPTOR was significantly outperformed by SPECTRE and COLIPRI [14] on both external non-oncologic chest CT datasets: CAPTOR reached an AUC of 0.796 (95%CI: 0.785-0.803) on CT-RATE vs. 0.847 (95%CI: 0.836-0.854) for COLIPRI and an AUC of 0.689 (95%CI: 0.669-0.706) on RAD-ChestCT [33] vs. 0.737 (95%CI: 0.719-0.752) for COLIPRI (Appendix B.7).

3.3 Survival prediction

CAPTOR outperformed all baseline models on three of the four prognostic tasks, achieving AUCs of 0.640 (95% CI: 0.567–0.676) on NSCLC-Radiogenomics [22], 0.637 (95% CI: 0.523–0.697) on CACK-KiTS [23], and 0.587 (95% CI: 0.546–0.609) on NSCLC-Radiomics [21]. On Colorectal-Liver-Metastases [24], CAPTOR ranked second with an AUC of 0.521 (95% CI: 0.427–0.604), behind COLIPRI [14] (0.549, 95%

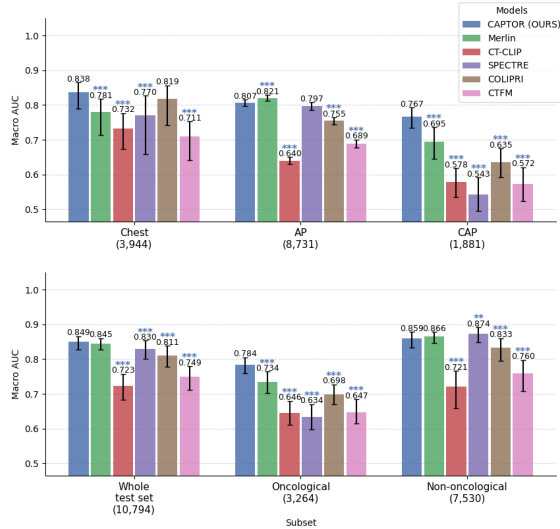


Figure 3: Bar plot showing the classification scores of CAPTOR and five state-of-the-art foundation models on our test set, stratified by visible anatomical region and clinical indication. The number of studies in each subset is in parentheses. One XGBoost was trained for each medical condition and model using the image embedding extracted with the model’s image encoder. For each subset and model, the macro-AUC is the unweighted mean of the AUC scores for each medical condition relevant to the subset among the 26 conditions studied. The confidence intervals and the pairwise comparison were obtained with a patient-level bootstrap and a patient-level permutation test respectively, both based on 2,000 iterations. AP = Abdomen-Pelvis, CAP=Chest-Abdomen-Pelvis. *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$

CI: 0.471–0.621) (Figure 5). Although few differences are statistically significant, the overall ranking pattern was consistent with the primary classification benchmark.

Computation times for the Colorectal-Liver-Metastases are presented in Appendix B.5.

3.4 Retrieval

CAPTOR achieved the highest text-image retrieval performance on CAP (Recall@10 of 47.8, 95% CI: 44.2–51.6), oncologic (83.4, 95% CI: 78.4–88.6), and chest (64.3, 95% CI: 51.9–84.0) subsets, while Merlin performed best on the whole test set (235, 95% CI: 205–261), abdomen-pelvis (238, 95% CI: 230–245), and non-oncologic subsets (236, 95% CI:

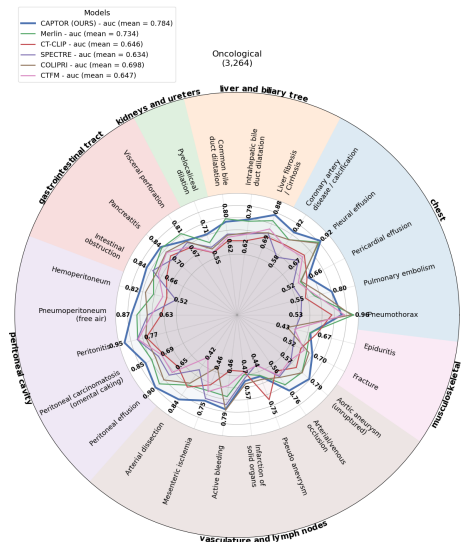


Figure 4: Per-condition classification performance of CAPTOR and five baselines on the oncological subset, for each of the 26 medical conditions studied. One XGBoost was trained for each medical condition and model, using the appropriate image embedding (i.e. the corresponding regional crop for CAPTOR and the corresponding series for the baselines) extracted with the model’s image encoder.

199–272). Image-image retrieval showed the same ranking, with CAPTOR achieving a MAP@10 of 1.30 (95%CI: 1.25–1.35) on the CAP subset, 1.41 (95%CI: 1.37–1.46) on the oncological subset, and 1.23 (95%CI: 1.20–1.27) on the chest subset (Tables 2).

3.5 Impact of training data volume

CAPTOR-100K, trained on the full pretraining dataset (internal and public image-report datasets), demonstrated comparable performance to CAPTOR on all subsets of the classification task, the biggest differences being for the CAP subset (0.744, 95%CI: 0.703–0.779 vs. 0.767, 95%CI: 0.734–0.793 for CAPTOR), and the oncological subset (0.771, 95%CI: 0.743–0.796 vs. 0.784, 95%CI: 0.759–0.805 for CAPTOR), although they were not significant. CAPTOR-100K statistically significantly outperformed CAPTOR on external non-oncologic chest CT datasets, reaching an AUC of 0.820 (95%CI: 0.813–0.827) on CT-RATE [11] and 0.719 (95%CI: 0.698–0.735) on RAD-ChestCT [33], while still outperformed by SPECTRE [13] and COLIPRI [14], however not statistically significantly (Appendix B.7).

Dataset	K	Retrieval text-image				Retrieval image-image				
		CAPTOR (OURS)	Merlin	CT-CLIP	COLIPRI	CAPTOR (OURS)	Merlin	CT-CLIP	COLIPRI	CTFM
Whole test set 10,794	1	<u>307</u> (266-348)	722 (622-817)	20.0 (8.88-31.1)	188 (104-281)	<u>1.78</u> (1.72-1.82)	1.82 (1.75-1.85)	1.35 (1.27-1.40)	1.63 (1.54-1.68)	1.40 (1.33-1.46)
	5	<u>179</u> (159-199)	353 (307-394)	12.2 (6.60-18.7)	98.4 (57.8-144)	<u>1.44</u> (1.39-1.45)	1.46 (1.41-1.46)	1.20 (1.12-1.20)	1.35 (1.28-1.36)	1.23 (1.16-1.24)
	10	<u>139</u> (125-153)	235 (205-261)	11.3 (6.71-16.6)	70.1 (42.6-101)	<u>1.46</u> (1.44-1.50)	1.48 (1.45-1.51)	1.20 (1.16-1.24)	1.35 (1.31-1.40)	1.23 (1.20-1.28)
	50	<u>65.8</u> (60.9-70.4)	77.1 (67.8-84.8)	8.02 (4.97-11.8)	30.7 (21.7-40.5)	<u>1.59</u> (1.57-1.64)	1.60 (1.58-1.65)	1.24 (1.21-1.31)	1.42 (1.39-1.49)	1.28 (1.26-1.35)
	100	<u>43.1</u> (40.2-45.8)	45.9 (40.9-50.2)	6.62 (4.24-9.54)	21.0 (16.0-26.5)	<u>1.61</u> (1.59-1.66)	1.63 (1.60-1.68)	1.24 (1.21-1.31)	1.43 (1.40-1.50)	1.29 (1.26-1.36)
Chest 3,944	1	<u>145</u> (111-197)	23.0 (14.0-36.1)	22.0 (10.9-30.4)	184 (111-231)	1.33 (1.28-1.37)	1.27 (1.22-1.31)	1.17 (1.11-1.22)	<u>1.29</u> (1.22-1.34)	1.18 (1.11-1.23)
	5	<u>85.2</u> (67.9-113)	19.0 (14.2-26.1)	12.8 (7.35-16.7)	92.0 (58.4-114)	1.22 (1.18-1.25)	1.18 (1.14-1.20)	1.11 (1.06-1.14)	<u>1.19</u> (1.14-1.22)	1.12 (1.07-1.14)
	10	64.3 (51.9-84.0)	15.3 (11.9-20.0)	11.6 (7.04-15.0)	63.3 (41.1-78.0)	1.23 (1.20-1.27)	<u>1.19</u> (1.16-1.23)	1.11 (1.08-1.16)	1.19 (1.15-1.23)	1.12 (1.08-1.16)
	50	28.6 (24.4-35.4)	9.10 (7.79-11.1)	8.08 (5.50-9.92)	25.7 (18.8-30.0)	1.26 (1.23-1.30)	<u>1.23</u> (1.20-1.28)	1.12 (1.08-1.18)	1.18 (1.14-1.24)	1.13 (1.09-1.18)
	100	18.8 (16.2-22.9)	6.71 (5.86-7.90)	6.58 (4.65-7.93)	16.8 (13.3-19.1)	1.25 (1.21-1.29)	<u>1.22</u> (1.18-1.27)	1.10 (1.06-1.16)	1.16 (1.11-1.22)	1.11 (1.06-1.16)
AP 8,731	1	<u>270</u> (241-305)	733 (685-782)	6.00 (2.00-11.0)	29.0 (19.0-39.9)	<u>1.72</u> (1.66-1.76)	1.78 (1.71-1.80)	1.25 (1.18-1.29)	1.53 (1.46-1.56)	1.28 (1.23-1.33)
	5	<u>157</u> (148-168)	357 (343-372)	4.40 (2.60-6.25)	22.0 (17.8-26.1)	<u>1.41</u> (1.36-1.42)	1.43 (1.38-1.44)	1.14 (1.07-1.14)	1.29 (1.23-1.29)	1.17 (1.10-1.17)
	10	<u>122</u> (116-129)	238 (230-245)	4.00 (2.80-5.23)	18.2 (15.6-20.7)	<u>1.43</u> (1.40-1.46)	1.45 (1.43-1.48)	1.14 (1.11-1.17)	1.30 (1.27-1.33)	1.17 (1.14-1.20)
	50	<u>56.6</u> (54.8-58.2)	78.0 (76.2-79.8)	2.92 (2.40-3.42)	12.8 (11.8-13.7)	<u>1.54</u> (1.53-1.59)	1.56 (1.54-1.61)	1.16 (1.15-1.21)	1.35 (1.34-1.40)	1.20 (1.19-1.25)
	100	<u>36.9</u> (36.0-37.8)	46.2 (45.3-47.1)	2.74 (2.41-3.07)	9.94 (9.36-10.5)	<u>1.56</u> (1.54-1.61)	1.59 (1.56-1.63)	1.17 (1.16-1.22)	1.36 (1.34-1.41)	1.20 (1.19-1.25)
CAP 1,881	1	109 (90.3-129)	<u>25.0</u> (15.9-34.9)	6.00 (1.99-11.2)	24.0 (14.6-34.3)	1.49 (1.40-1.56)	<u>1.37</u> (1.30-1.46)	1.24 (1.15-1.32)	1.36 (1.27-1.43)	1.22 (1.13-1.31)
	5	62.4 (56.1-68.9)	<u>19.0</u> (15.3-22.6)	3.60 (2.02-5.44)	13.6 (10.5-17.0)	1.28 (1.22-1.32)	<u>1.22</u> (1.16-1.26)	1.13 (1.06-1.16)	1.21 (1.14-1.24)	1.12 (1.05-1.16)
	10	47.8 (44.2-51.6)	<u>15.1</u> (12.8-17.4)	3.30 (2.24-4.54)	10.5 (8.55-12.5)	1.30 (1.25-1.35)	<u>1.24</u> (1.19-1.29)	1.13 (1.09-1.19)	1.20 (1.16-1.26)	1.13 (1.09-1.19)
	50	19.8 (18.9-20.6)	<u>8.14</u> (7.42-8.86)	2.48 (2.04-2.94)	6.98 (6.32-7.62)	1.33 (1.29-1.40)	<u>1.27</u> (1.23-1.34)	1.12 (1.08-1.20)	1.19 (1.15-1.26)	1.14 (1.10-1.21)
	100	12.6 (12.2-13.0)	<u>5.62</u> (5.20-5.99)	2.14 (1.85-2.44)	5.16 (4.79-5.56)	1.33 (1.28-1.39)	<u>1.27</u> (1.23-1.34)	1.12 (1.07-1.19)	1.18 (1.13-1.25)	1.12 (1.08-1.20)
Oncological 3,264	1	224 (195-253)	38.0 (26.1-51.2)	6.00 (1.99-11.0)	<u>41.0</u> (29.1-53.9)	1.65 (1.56-1.70)	<u>1.47</u> (1.40-1.55)	1.29 (1.20-1.36)	1.37 (1.29-1.44)	1.29 (1.21-1.36)
	5	115 (106-125)	<u>27.6</u> (23.1-32.2)	6.20 (4.19-8.43)	22.8 (18.7-27.1)	1.40 (1.34-1.42)	<u>1.29</u> (1.23-1.32)	1.17 (1.10-1.19)	1.23 (1.16-1.25)	1.19 (1.11-1.21)
	10	83.4 (78.4-88.6)	<u>21.5</u> (18.8-24.2)	4.60 (3.32-5.90)	16.9 (14.4-19.5)	1.41 (1.37-1.46)	<u>1.30</u> (1.26-1.35)	1.17 (1.12-1.22)	1.23 (1.18-1.27)	1.19 (1.14-1.24)
	50	33.5 (32.4-34.6)	<u>11.2</u> (10.3-12.0)	3.28 (2.79-3.78)	9.94 (9.16-10.7)	1.48 (1.45-1.55)	<u>1.38</u> (1.34-1.44)	1.18 (1.15-1.25)	1.23 (1.20-1.30)	1.21 (1.18-1.27)
	100	21.0 (20.5-21.5)	<u>8.09</u> (7.61-8.60)	2.90 (2.56-3.23)	7.38 (6.93-7.88)	1.49 (1.45-1.55)	<u>1.38</u> (1.34-1.44)	1.17 (1.13-1.24)	1.21 (1.17-1.28)	1.20 (1.16-1.26)
Non-oncological 7,530	1	190 (160-224)	745 (619-880)	22.0 (7.78-36.1)	<u>191</u> (93.7-288)	<u>1.69</u> (1.61-1.74)	1.80 (1.72-1.83)	1.27 (1.17-1.35)	1.60 (1.51-1.65)	1.33 (1.23-1.40)
	5	<u>116</u> (102-132)	365 (308-423)	14.4 (6.30-22.3)	99.0 (53.6-142)	<u>1.38</u> (1.31-1.39)	1.44 (1.38-1.45)	1.15 (1.06-1.17)	1.32 (1.25-1.34)	1.18 (1.10-1.20)
	10	<u>90.3</u> (80.5-101)	236 (199-272)	11.9 (5.55-18.2)	69.3 (38.8-98.1)	<u>1.41</u> (1.37-1.45)	1.47 (1.44-1.50)	1.16 (1.11-1.21)	1.35 (1.30-1.39)	1.19 (1.14-1.25)
	50	<u>43.0</u> (39.7-46.7)	74.2 (64.0-84.8)	8.24 (4.10-12.4)	27.6 (18.7-36.5)	<u>1.52</u> (1.48-1.58)	1.58 (1.55-1.63)	1.18 (1.14-1.27)	1.41 (1.36-1.48)	1.23 (1.18-1.31)
	100	<u>28.7</u> (27.0-30.6)	42.8 (37.2-48.5)	6.41 (3.27-9.48)	18.7 (13.6-23.5)	<u>1.54</u> (1.49-1.60)	1.59 (1.55-1.64)	1.19 (1.14-1.27)	1.42 (1.36-1.49)	1.23 (1.18-1.31)

Table 2: Zero-shot text-image (Recall@K) and image-image (MAP@K) retrieval performance of CAPTOR and baseline models, expressed as fold-increase over the random baseline (best in bold and second best underlined) with 95% confidence intervals in parentheses, by anatomical and oncologic subset. K is the number of retrieved images. SPECTRE [13] could not be evaluated (unavailable projection-layer weights) and CTFM [17], a vision model, was excluded from text-image retrieval.

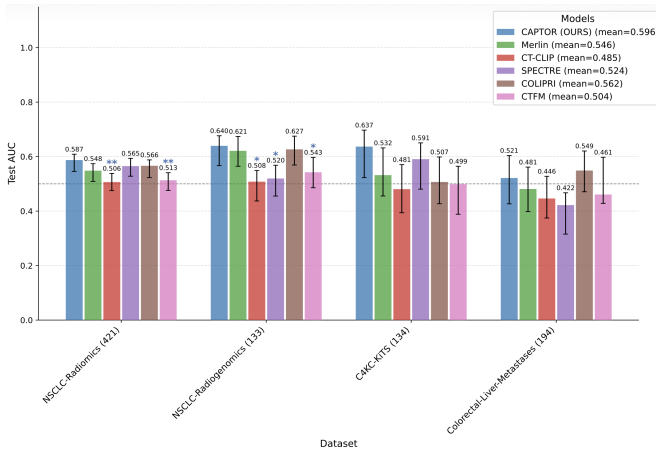


Figure 5: Two-year survival-prediction performance (AUC) of CAPTOR and five baseline models on four oncologic TumorImagingBench1 cohorts (NSCLC-Radiomics21, NSCLC-Radiogenomics22, C4KC-KiTS23, Colorectal-Liver-Metastases24), using kNN classifiers on frozen image embeddings, averaged over 10 random splits. Statistical significance was assessed using a permutation test (2,000 permutations). The number of studies for each dataset is indicated between parentheses. ** = $p < 0.01$, * = $p < 0.05$

CAPTOR outperformed CAPTOR-100K on the two abdomen datasets of the four oncological test cohorts while CAPTOR-100K demonstrated superior performance on the two chest CT datasets (CAPTOR-100K AUC : 0.616, 95% CI: 0.581-0.642 on NSCLC-Radiomics [21] and 0.652, 95% CI: 0.583-0.689 on NSCLC-Radiogenomics [22]), though none of the differences between the models were statistically significant.

For retrieval, CAPTOR-100K outperformed all baseline models on the whole test set, Chest, AP and non-oncological subsets, reaching a Recall@10 for the whole test set of 310 (95% CI:302–319) and MAP@10 of 1.50 (95% CI: 1.47-1.54), but was outperformed by CAPTOR on the CAP and oncologic subsets. Table 3 summarizes the performance differences between CAPTOR and CAPTOR-100K across all downstream tasks. Detailed figures of CAPTOR-100K performances compared to baselines are in Appendix B.6

We further evaluated the effect of training data volume on classifier performance by varying the fraction of the training set used during XGBoost training (Figure 6). Increasing the amount of training data generally leads to improved performance, although the magnitude and consistency of these gains vary

Datasets Task	Subset	CAPTOR	CAPTOR-100K
Global test set Medical condition classification	Whole	0.849 (0.828–0.866)	0.846 (0.824–0.862)
	Chest	0.838 (0.789–0.865)	0.828 (0.778–0.860)
	AP	0.807 (0.797–0.816)	0.808 (0.798–0.818)
	CAP	0.767 (0.734–0.793)	0.744 (0.703–0.779)
	Oncological	0.784 (0.759–0.805)	0.771 (0.743–0.796)
	Non-oncological	0.859 (0.833–0.879)	0.856 (0.831–0.874)
Oncological test sets Two-year survival prediction	NSCLC-Radiomics	0.587 (0.546–0.609)	0.616 (0.581–0.642)
	NSCLC-Radiogenomics	0.640 (0.567–0.676)	0.652 (0.583–0.689)
	C4KC-KiTS	0.637 (0.523–0.697)	0.556 (0.447–0.633)
	Colorectal-Liver Metastases	0.521 (0.427–0.604)	0.482 (0.403–0.576)
Global test set Text-Image retrieval	Whole	139 (125–153)	310 (302–319)
	Chest	64.3 (51.9–84.0)	95.2 (86.2–101)
	AP	122 (116–129)	254 (245–262)
	CAP	47.8 (44.2–51.6)	38.6 (35.3–41.9)
	Oncological	83.4 (78.4–88.6)	71.7 (67.3–76.6)
	Non-oncological	90.3 (80.5–101)	265 (255–276)
Global test set Image-Image retrieval	Whole	1.46 (1.44–1.50)	1.50 (1.47–1.54)
	Chest	1.23 (1.20–1.27)	1.24 (1.21–1.28)
	AP	1.43 (1.40–1.46)	1.46 (1.44–1.49)
	CAP	1.30 (1.25–1.35)	1.28 (1.23–1.33)
	Oncological	1.41 (1.37–1.46)	1.39 (1.35–1.44)
	Non-oncological	1.41 (1.37–1.45)	1.47 (1.42–1.51)

Table 3: Effect of training-data volume on CAPTOR performance. Comparison of CAPTOR and CAPTOR-100K performances across all downstream tasks with 95%confidence intervals in parentheses. Retrieval results are shown for K=10, with K the number of retrieved images considered.

across datasets. CAPTOR consistently achieves the highest mean AUC across all training fractions compared to baselines on the whole test set, the CAP and the oncological subsets, demonstrating strong robustness in low-data settings.

4 Discussion

In this multicenter study comparing six CT vision-language foundation models across classification, prognostic benchmarks and retrieval, CAPTOR achieved the highest performance on CAP and oncologic CT subsets. The principal finding of this study is that a foundation model trained on clinically representative oncologic CAP CT examinations achieved superior performance in the intended deployment setting compared with both region-specific and generalist CT foundation models, supporting the hypothesis that alignment between training data and clinical use case is a major determinant of foundation-model performance.

A key architectural contribution enabling this performance is its region-centered preprocessing strategy, in which TotalSegmentator [28] masks were used to automatically crop anatomical sub-regions prior

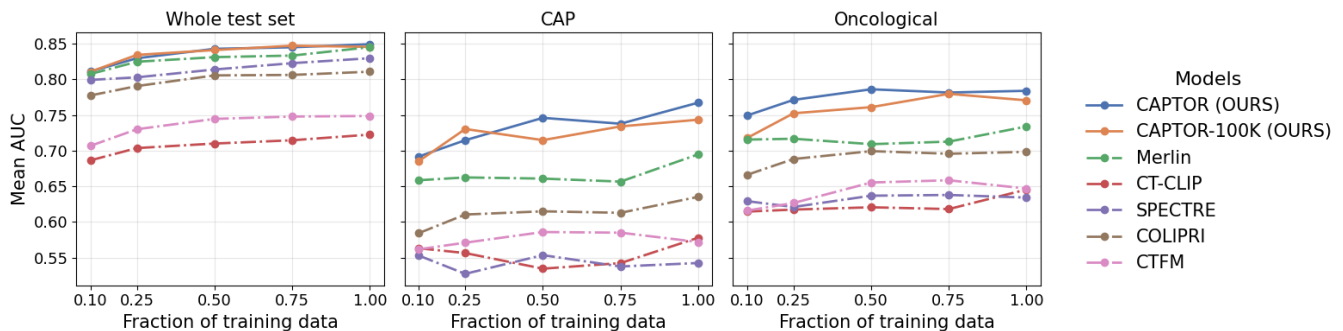


Figure 6: Classification performance (macro-AUC) as a function of labeled training-set fraction (10-100%) for CAPTOR, CAPTOR-100K, and five baseline models, on (left) the whole test set, (middle) the CAP subset, and (right) the oncologic subset.

to encoding, allowing CAPTOR to generate independent embeddings for each sub-region, preserving region-specific representations while enabling full CAP coverage.

This allowed CAPTOR to process full CAP volumes within the same framework as region-specific models such as Merlin [12], without sacrificing the whole-examination context that oncologic interpretation requires, likely contributing to CAPTOR’s advantage on CAP and oncologic subsets.

An important finding was that CAPTOR achieved performance comparable to, and in several analyses slightly exceeding, CAPTOR-100K despite being trained on substantially fewer examinations. CAPTOR performed best on the intended oncologic CAP population, whereas CAPTOR-100K showed advantages on non-oncologic external datasets and retrieval. These findings indicate that distribution differences between pretraining and downstream populations (data drift) also matter for foundation models, highlighting a trade-off between specialization and generalizability rather than a simple scaling effect.

The retrieval experiments provide complementary evidence that the improvements observed in classification are attributable to stronger image representations rather than solely to downstream classifier optimization. CAPTOR demonstrated consistently strong image-image and text-image retrieval performance, particularly on oncologic and CAP examinations, indicating improved cross-modal alignment between CT images and radiology reports.

This study has several limitations. First, this is a retrospective study, with no prospective validation. Second, reference labels were automatically extracted from radiology reports using a large language model rather than manual expert annotation. Extraction

performance was evaluated on the public dataset CT-RATE [11], and a random sample of 25 reports was manually reviewed to assess extraction quality, although, omissions in the original reports and LLM errors may have introduced some degree of label noise at scale. Third, CAPTOR is a case-level foundation model: it aggregates information across large anatomical regions into a single study-level embedding, enabling efficient inference and strong examination-level performance, but is not designed for lesion-level localization or characterization. Architectures such as SPECTRE [13], which preserve fine-grained spatial information, remain better suited to lesion-level tasks, and our results should be interpreted within the context of global examination understanding rather than lesion analysis. Fourth, several external oncologic datasets had small sample sizes, limiting statistical power for pairwise comparisons; although CAPTOR generally outperformed competing methods on these datasets, larger multicenter cohorts are needed to confirm these findings. Finally, CAPTOR’s generalizability beyond its intended oncologic CAP CT setting remains limited: on the non-oncologic CT-RATE and RAD-Chest CT [33] datasets, it was outperformed by generalist models such as SPECTRE and COLIPRI [14], consistent with its oncology-enriched training distribution. While CAPTOR remained competitive despite this domain shift, specialization for a targeted clinical use case may come at the expense of performance on substantially different imaging populations, and future work should explore strategies that better balance specialization and generalizability.

In conclusion, CAPTOR demonstrates that a vision-language foundation model trained on a clinically targeted, oncology-enriched CAP CT dataset

can outperform both region-specific and generalist models on its intended use case, without requiring large-scale heterogeneous data. These results suggest that dataset composition and clinical alignment may be at least as important as scale in CT foundation model development. To support reproducibility and community adoption, the training code and model weights will be made publicly available. Future work should explore architectures and training strategies that preserve specialization gains while extending generalizability across broader clinical use cases.

5 Acknowledgments

The authors used OpenAI’s ChatGPT (version 5) and Claude Sonnet 4.6 to assist with grammar and language when writing this article. This work was performed using HPC resources from GENCI–IDRIS (2026-103370).

References

- [1] Global Cancer Observatory. Estimated number of new cases from 2022 to 2045, Both sexes, age [0-85+], February 2024. URL <https://gco.iarc.who.int/tomorrow/en/dataviz/isotype>.
- [2] Kenneth D. Hopper, Kishor Singapuri, and Arkady Finkel. Body CT and Oncologic Imaging. *Radiology*, 2000. doi: 10.1148/radiology.215.1.r00ap1727. URL <https://sci-hub.fr/10.1148/radiology.215.1.r00ap1727>.
- [3] ASCO GUIDELINES Bundle - Lung Cancer Surveillance After Definitive Curative-Intent Therapy, . URL <https://eguideline.guidelinecentral.com/i/1475471-lung-cancer-surveillance-after-definitive-curative-intent-therapy/1?>
- [4] A. Zer, M.-J. Ahn, F. Barlesi, L. Bubendorf, D. De Ruysscher, P. Garrido, O. Gautschi, L.E. Hendriks, P.A. Jänne, K.M. Kerr, C. Mascaux, T. Mitsudomi, S. Peters, C. Rolfo, A. Sacher, S. Senan, P. Ugalde, and N.B. Leighl. Early and locally advanced non-small-cell lung cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Annals of Oncology*, 36(11):1245–1262, November 2025. ISSN 09237534. doi: 10.1016/j.annonc.2025.08.003. URL <https://linkinghub.elsevier.com/retrieve/pii/S0923753425009238>.
- [5] R. Glynne-Jones, L. Wyrwicz, E. Tiret, G. Brown, C. Rödel, A. Cervantes, and D. Arnold. Rectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. *Annals of Oncology*, 28:iv22–iv40, July 2017. ISSN 0923-7534. doi: 10.1093/annonc/mdx224. URL <https://www.sciencedirect.com/science/article/pii/S0923753419421522>.
- [6] R. J. M. Bruls and R. M. Kwee. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights into Imaging*, 11:121, November 2020. ISSN 1869-4101. doi: 10.1186/s13244-020-00925-z. URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC7683675/>.
- [7] Mina S. Makary, MD and Noah Takacs, BS. Are We Prepared for a Looming Radiologist Shortage? URL <https://www.diagnosticimaging.com/view/are-we-prepared-for-a-looming-radiologist-shortage->.
- [8] Anika G. Patel, Victor J. Pizzitola, C. Daniel Johnson, Nan Zhang, and Maitray D. Patel. Radiologists Make More Errors Interpreting Off-Hours Body CT Studies during Overnight Assignments as Compared with Daytime Assignments. *Radiology*, 297(2):374–379, November 2020. ISSN 0033-8419. doi: 10.1148/radiol.2020201558. URL <https://pubs.rsna.org/doi/10.1148/radiol.2020201558>.
- [9] Emanuele Neri, Nandita de Souza, Adrian Brady, Angel Alberich Bayarri, Christoph D. Becker, Francesca Coppola, Jacob Visser, and European Society of Radiology (ESR). What the radiologist should know about artificial intelligence – an ESR white paper. *Insights into Imaging*, 10(1):44, April 2019. ISSN 1869-4101. doi: 10.1186/s13244-019-0738-2. URL <https://doi.org/10.1186/s13244-019-0738-2>.
- [10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In

Proceedings of the 38th International Conference on Machine Learning, pages 8748–8763. PMLR, July 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.

- [11] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, February 2026. ISSN 2157-846X. doi: 10.1038/s41551-025-01599-y. URL <https://www.nature.com/articles/s41551-025-01599-y>.
- [12] Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Robbie Holland, Cesar Truys, Christian Bluethgen, Yufu Wu, Long Lian, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro Junior, Neera Ahuja, Jason Fries, Nigam H. Shah, Greg Zaharchuk, Marc Willis, Adam Yala, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, and Akshay S. Chaudhari. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, pages 1–11, March 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10181-8. URL <https://www.nature.com/articles/s41586-026-10181-8>.
- [13] Cris Claessens, Christiaan Viviers, Giacomo D’Amicantonio, Egor Bondarev, and Fons van der Sommen. Scaling Self-Supervised and Cross-Modal Pretraining for Volumetric CT Transformers, November 2025. URL <http://arxiv.org/abs/2511.17209>. arXiv:2511.17209 [cs].
- [14] Tassilo Wald, Ibrahim Ethem Hamamci, Yuan Gao, Sam Bond-Taylor, Harshita Sharma, Maximilian Ilse, Cynthia Lo, Olesya Melnichenko, Anton Schwaighofer, Noel C. F. Codella, Maria Teodora Wetscherek, Klaus H. Maier-Hein, Panagiotis Korfiatis, Valentina Salvatelli, Javier Alvarez-Valle, and Fernando Pérez-García. Comprehensive language-image pre-training for 3D medical image understanding, January 2026. URL <http://arxiv.org/abs/2510.15042>. arXiv:2510.15042 [cs].
- [15] Jing Li, Jing Fu, Xin Geng, and Hongyang Wang. Foundation models in clinical oncology: Progresses and perspectives. *Cancer Letters*, 639:218220, February 2026. ISSN 0304-3835. doi: 10.1016/j.canlet.2025.218220. URL <https://www.sciencedirect.com/science/article/pii/S030438352500792X>.
- [16] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Kun-Hsing Yu, Sierra Willens, Francesca Maria Olguin, Jeffrey J. Nirschl, Joel Neal, Maximilian Diehn, Sen Yang, and Ruijiang Li. A vision–language foundation model for precision oncology. *Nature*, 638 (8051):769–778, February 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08378-w. URL <https://www.nature.com/articles/s41586-024-08378-w>.
- [17] Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressen, Benjamin H. Kann, Andriy Fedorov, Raymond H. Mak, and Hugo J. W. L. Aerts. Vision Foundation Models for Computed Tomography, 2025. URL <https://arxiv.org/abs/2501.09001>. Version Number: 2.
- [18] Hugo Aerts, Suraj Pai, Ibrahim Hadzic, Andrey Fedorov, and Raymond Mak. Foundation model embeddings for quantitative tumor imaging biomarkers, May 2025. ISSN: 2693-5015.
- [19] Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P Lungren, Curtis P Langlotz, Serena Yeung, Nigam H Shah, and Jason A Fries. INSPECT: A Multimodal

- Dataset for Pulmonary Embolism Diagnosis and Prognosis.
- [20] Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, and Zhiwei Xiong. BIMCV-R: A Landmark Dataset for 3D CT Text-Image Retrieval, 2024. URL <https://arxiv.org/abs/2403.15992>. Version Number: 2.
 - [21] NSCLC-RADIOMICS, . URL <https://www.cancerimagingarchive.net/collection/nsclc-radiomics/>.
 - [22] NSCLC-RADIOGENOMICS, . URL <https://www.cancerimagingarchive.net/collection/nsclc-radiogenomics/>.
 - [23] C4KC-KITS, . URL <https://www.cancerimagingarchive.net/collection/c4kc-kits/>.
 - [24] COLORECTAL-LIVER-METASTASES, . URL <https://www.cancerimagingarchive.net/collection/colorectal-liver-metastases/>.
 - [25] Mistral Large 2.1 - Mistral AI, . URL <https://docs.mistral.ai/models/mistral-large-2-1-24-11>.
 - [26] Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences, January 2022. URL <https://arxiv.org/abs/2201.11838v3>.
 - [27] João Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, July 2017. doi: 10.1109/CVPR.2017.502. URL <https://ieeexplore.ieee.org/document/8099985>. ISSN: 1063-6919.
 - [28] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence*, 5(5):e230024, September 2023. ISSN 2638-6100. doi: 10.1148/ryai.230024. URL <http://pubs.rsna.org/doi/10.1148/ryai.230024>.
 - [29] Adam E. Flanders, Xindi Wang, Carol C. Wu, Felipe C. Kitamura, George Shih, John Mongan, and Yifan Peng. The Evolution of Radiology Image Annotation in the Era of Large Language Models. *Radiology: Artificial Intelligence*, 7(4):e240631, July 2025. doi: 10.1148/ryai.240631. URL <https://pubs.rsna.org/doi/10.1148/ryai.240631>.
 - [30] Felix J. Dorfner, Liv Jürgensen, Leonhard Donle, Fares Al Mohamad, Tobias R. Bodenmann, Mason C. Cleveland, Felix Busch, Lisa C. Adams, James Sato, Thomas Schultz, Albert E. Kim, Jameson Merkow, Keno K. Bressen, and Christopher P. Bridge. Comparing Commercial and Open-Source Large Language Models for Labeling Chest Radiograph Reports. *Radiology*, 313(1):e241139, October 2024. ISSN 0033-8419. doi: 10.1148/radiol.241139. URL <https://pubs.rsna.org/doi/10.1148/radiol.241139>.
 - [31] Introducing GPT-5.4, July 2026. URL <https://openai.com/index/introducing-gpt-5-4/>.
 - [32] GPT-4o mini: advancing cost-efficient intelligence, . URL <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
 - [33] Rachel Lea Draelos, David Dov, Maciej A Mazurowski, Joseph Y. Lo, Ricardo Henao, Geoffrey D. Rubin, and Lawrence Carin. RAD-ChestCT Dataset, October 2020. URL <https://zenodo.org/record/6406114>.
 - [34] Tetiana Manzhos, Tetiana Ianevych, and Olga Melnyk. Average precision at cutoff k under random rankings: expectation and variance. *Modern Stochastics: Theory and Applications*, pages 1–18, April 2026. ISSN 2351-6046, 2351-6054. doi: 10.15559/26-VMSTA298. URL <https://www.vmsta.org/journal/VMSTA/article/353>.
 - [35] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting Transformer-based Language Models to Radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, July 2022. doi: 10.1148/ryai.210258. URL <https://pubs.rsna.org/doi/10.1148/ryai.210258>.

A Supplementary methods

A.1 Internal dataset

For each study, series were selected via a priority rule favoring full-coverage and contrast-enhanced acquisitions. The series selection priority is the following: CAP > Chest + AP > CA + Pelvis > Chest + Abdomen + Pelvis > Unknown > AP > CA > Chest or Abdomen or Pelvis. If several series correspond to a category (several CAP series for example), we select the series based on the phase and kernel type : Venous + Smooth > Unenhanced + Smooth > Delayed + Smooth > Arterial + Smooth > Unknown + Sharp > Unknown + Smooth > Venous + Sharp > First available series.

A.2 External dataset series selection criteria

INSPECT was excluded from the testing set because the reports only include the Impressions section which is not detailed enough to extract reliable labels, and BIMCVR was excluded from the test set due to its limited size and lack of descriptive information.

A.3 More information on oncological datasets

The four public oncological datasets we used among those in TumorImagingBench are the following :

- NSCLC-Radiomics [21] : CT scans from 421 non-small cell lung cancer (NSCLC) patients, who received radiation therapy or chemo-radiation therapy. Pretreatment CT scans, expert segmentation of Gross Tumor Volume (GTV) and clinical outcome data are available for each patient. The median survival time of this study is 17.9 months.

- NSCLC-Radiogenomics [22] : Cohort of 211 NSCLC patients with preoperative CT and PET/CT scans. We used images from the 133 patients with GTV annotated like TumorImagingBench. The median survival time of these patients is 44.5 months.

- C4KC-KiTS [23] : CT scans and tumor and kidney segmentations from renal tumor patients with partial or radical nephrectomy. Of the 210 patients in the dataset, 134 remain after filtering for survival data at 2 years. The median survival time of these patients is 15.5 months.

- Colorectal-Liver-Metastases [24] : Preoperative CT scans and liver/tumor/vessels segmentation from

194 patients who underwent colorectal liver metastases (CRLM) resection. The median survival time of this study is 44.5 months.

We did not use the two other datasets originally present in TumorImagingBench because they are lesion-level, which is not our aim with CAPTOR.

A.4 More information on preprocessing

CT volumes were reformatted to RAS orientation, resampled to isotropic in-plane resolution (1.5 mm) and 3-mm slice spacing, intensity-normalized, padded, and cropped to a fixed input size of 224x224x160 voxels. The full list of regions we used is the same as Merlin: chest, liver and biliary tree, gallbladder, spleen, pancreas, adrenal glands, kidneys and ureters, gastrointestinal tract, peritoneal cavity, pelvic organs, vasculature and lymph nodes, musculoskeletal

Regarding report preprocessing, reports were automatically checked to make sure they did not contain unreadable characters. 94% of reports were acquired already in English (originally in English or already translated); a smaller subset (1,805 studies, 6% of the dataset) remained in Spanish and was used as-is for training, without separate handling. In the test set, Spanish reports represented a smaller share (2%) and were all excluded from all analyses.

A.5 Training hyperparameters

We used an Adam optimizer with 1.25e-6 learning rate, $\beta = (0.9, 0.99)$ and used gradient checkpointing for both the visual and text encoders. We used a batch size of 18 on a single H100 GPU.

A.6 Baseline Models

We compared our model with five state-of-the-art foundation models. Two of them were trained on region-specific CT datasets:

- MERLIN [12]: 3D contrastive vision-language model based on an inflated ResNet-152 image encoder and a Clinical Longformer text encoder. It was trained on approximately 25,000 abdomen-pelvis CT-report pairs from the STARR repository (Merlin dataset), consisting predominantly of non-oncologic examinations.

- CT-CLIP [11]: a 3D contrastive vision-language model using transformer vision and text encoders (adapted from CT-ViT for the vision encoder and

CXR-Bert for the text encoder), trained on the CT-RATE dataset comprising approximately 25,000 non-oncologic chest CT-report pairs.

The others are trained on broader datasets:

- SPECTRE [13]: a 3D Vision Transformer (ViT) pretrained using a combination of self-supervised and contrastive vision-language objectives. It was trained on 229,619 CT series from multiple public datasets spanning chest, abdomen, pelvis, and whole-body imaging (including CT-RATE, MERLIN, INSPECT, NLST, AMOS, AbdomenAtlas, PANORAMA, and AbdomenCT-1K). Although anatomically diverse, the training corpus was largely non-oncologic, with pancreatic cancer data (PANORAMA) representing approximately 1% of the total dataset.

- COLIPRI [14]: a transformer-based multi-modal foundation model combining contrastive vision-language and vision-only pretraining. CT training used the same public CT-report datasets included in CAPTOR-100K, complemented by large image-only MRI datasets (brain, knee/thigh, and whole-body MRI), making it the broadest multimodal medical foundation model evaluated.

- CT-FM [17]: a vision-only 3D foundation model trained using contrastive self-supervised learning on more than 148,000 CT examinations acquired across multiple cancer types and associated comorbidities. Unlike the other vision-language models, CT-FM does not use radiology reports during pretraining and learns image representations exclusively from CT volumes.

Table 4 summarizes their characteristics.

Model	VLM	Body Parts	Training Dataset(s)	Dataset size
CAPTOR (OURS)	✓	CAP	Internal	30K
Merlin	✓	Abdomen & Pelvis	Merlin	25K
CT-CLIP	✓	Chest	CT-RATE	50K volumes 25K scans
SPECTRE	✓	CAP	Multiple open CT datasets incl. CT-RATE, INSPECT, Merlin	230K
COLIPRI	✓	Chest	CT-RATE, NLST	100K
CT-FM	✗	CAP + others	Imaging Data Commons	148K

Table 4: Characteristics description of CAPTOR and the five baseline models we used in our analyses. VLM = Vision-Language Model

To compare the embedding quality of each model, we used the official documentation and code to preprocess the data and extract embeddings. For CAPTOR and the Merlin model we used the embeddings from the second to last layer, as it is the official image embedding method.

A.7 Label extraction - Prompt and conditions

We used zero-shot prompting and intentionally avoided few-shot prompting because selecting representative demonstration examples introduces an additional source of variability and may reduce reproducibility across datasets. The prompt we used for the label extraction with Mistral is the following:

System: "You are a radiologist reviewing radiology reports and classifying whether the patient has a condition or not."

Prompt: "Here is a radiology report from what radiologists can see in CT scans: {report_content} and here is a list of medical conditions that can potentially be stated in these reports: {list_conditions} You are to determine if:

- Each condition is stated or explained as detected in the report. In this case you will put the string 'True' in the field 'detected' of the right condition, the string 'False' otherwise.

- Each condition is stated or explained as not detected in the report. If the report states that there is a history of the condition in the patient, that counts as not detected condition. If the condition is stated as not detected in the report you will put the string 'True' in the field 'not_detected' in the field of the right condition, the string 'False' otherwise

- Each condition is not stated or explained at all in the report. In this case, you will put the string 'True' in the field 'not_mentioned' of the field of the right condition, the string 'False' otherwise. You will give the results as a JSON with the following schema {json_schema}, with strings 'True' or 'False' in the fields 'detected', 'not_detected', and 'not_mentioned' for each condition.

If the condition is stated as detected but the report also states that there is a history of this condition in the patient, please fill 'False' in the 'not_detected' field because a condition cannot be present and not present at the same time. There can only be one 'True' for each condition. Please do not add any condition that is not in the list I gave you."

The full list of conditions is the following : Arterial dissection, Hemoperitoneum, Mesenteric ischemia, Active bleeding, Infarction of solid organs, Pseudo aneurysm, Visceral perforation, Pneumoperitoneum (free air), Peritonitis, Pyelocaliceal dilation, Arterial/venous occlusion, Fracture, Pancreatitis, Aortic aneurysm (unruptured), Hemothorax, Intestinal obstruction, Liver fibrosis / Cirrhosis, Peritoneal car-

cinomatosis (omental caking), Pneumothorax, Pulmonary embolism, Pericardial effusion, Epiduritis, Pleural effusion, Peritoneal effusion, Intrahepatic bile duct dilatation, Common bile duct dilatation, Coronary artery disease calcification

A.8 XGBoost parameters and details on classification method

XGBoost was chosen as a strong non-linear classifier while using identical hyperparameters across all conditions and foundation models, ensuring that performance differences primarily reflected embedding quality rather than downstream model optimization. To train the XGBoost classifiers the binary logistic objective function (objective = binary:logistic) was used for classification. To account for the unequal class distribution, the positive class was weighted using the scale_pos_weight parameter. A moderate learning rate (learning_rate = 0.1) was selected, and model complexity was constrained by limiting tree depth (max_depth = 3) and by applying both L1 (reg_alpha = 0.5) and L2 (reg_lambda = 5) regularization. Additional variance reduction was introduced through row subsampling (subsample = 0.9) and feature subsampling (colsample_bytree = 0.9). The number of boosting iterations was fixed at 100 trees (n_estimators = 100). Overall, these settings were chosen to favor model generalization and robustness while limiting overfitting, particularly in datasets containing a limited number of positive samples.

For the baseline models we also used all available series, using each model’s method to get the embeddings from each series and using the embeddings from the right series for each condition (e.g. embeddings from the Chest series for chest conditions and from the AP series for AP conditions).

A.9 Further explanation on generalization evaluation

We tested all models on broadly used chest CT datasets: CT-RATE [11] and RAD-ChestCT [33].

The original RAD-Chest CT dataset contains 35,747 non-contrast chest CT scans from 19,661 adult patients of the Duke University Health System between January 2012 and April 2017. The publicly available version contains 3,630 chest CT scans, approximately 10% of the dataset. Each CT scan is annotated with a matrix of 84 abnormality labels x

52 location labels. For our use case we only used the 84 abnormality labels: if the abnormality is present in at least one location, then we set it to present, absent otherwise.

For both datasets we used the given labels, extracted from the reports with a fine-tuned RadBERT-RoBERTa-4m [35] for CT-RATE and with a rule-based method for RAD-ChestCT, and the given training and testing sets to train and test the classifiers.

A.10 Further explanation on retrieval metrics

For text-image retrieval, the random baseline for Retrieval@K is simply KN, where N is the total number of images, as each image has an equal probability of appearing in the top K retrieved images under random ranking. Consequently, both the random baseline and model scores increase with K, since the probability of retrieving the correct image within the top K candidates increases as K grows.

For image-image retrieval, following standard practice, the relevancy of an image with the reference image is computed with the Intersection over Union of labels, also known as the Jaccard index, thresholded at 0. The IoU consists in the number of positive labels in common (size of the intersection of labels) over the number of positive labels in total for both images (size of the union of labels). In short, in our computation, an image is considered relevant with another if they have the presence of at least one medical condition (positive label) in common. The MAP@K metric is then computed following the standard formulation [34]:

$$AP@K = \frac{1}{\min(R, K)} \sum_{i=1}^K P@i \cdot \mathbf{1}[rel(i)] \quad (1)$$

where R is the total number of relevant items among all the images for the query, P@i is the precision at rank i (the proportion of relevant images in the top i retrieved images), and $\mathbf{1}[rel(i)]$ indicates whether the item retrieved at rank i is relevant. The MAP@K is then the mean of AP@K over all queries. Importantly, only abnormal volumes (i.e., volumes presenting at least one positive label) are included as queries, as normal volumes have no relevant counterpart by construction under the Jaccard threshold of 0. The random image-image retrieval baseline is computed empirically via Monte Carlo simulation (50

trials, 500 queries per trial), where items are retrieved randomly. The random baseline increases from $K=1$ to $K=5$ before decreasing for larger K values, which is expected under the standard $AP@K$ formulation. As K increases, the normalization by $\min(R,K)$ can initially inflate scores for queries with few relevant items (if the single relevant item is retrieved within K positions, $AP@K$ equals 1 regardless of K) before the dilution of non-relevant items at higher ranks drives the score down. The baseline also varies across dataset subsets, reflecting differences in label density: subsets with more frequent or overlapping pathologies (e.g., Chest) exhibit higher random baselines than sparser subsets (e.g., AP), as a greater proportion of the gallery shares at least one label with any given query. Model performance is consequently expected to follow the same decreasing trend with K , since retrieving relevant items becomes relatively harder as the retrieved set grows.

B Complementary results

B.1 Detailed internal dataset characteristics

Table 5 presents the internal dataset detailed characteristics by country.

B.2 Mistral performance

Figure 7 displays the classification performance of Mistral Large 3 [25], GPT 5.4 [31] and GPT 4o mini [32] on the CT-RATE [11] dataset, using the dataset’s original labels extracted with a fine-tuned RadBERT-RoBERTa-4m [35] on 1000 reports.

B.3 Label proportions

Table 6 shows the proportions of each medical condition in each subset of our whole testing set (internal dataset combined with public datasets)

B.4 Statistical significance results by condition

Table 7 presents the results of the number of conditions for which the difference between our model CAPTOR and each baseline is statistically significant.

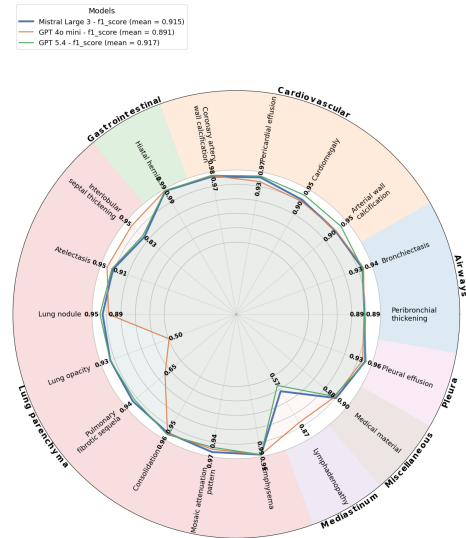


Figure 7: Radar plot comparing the per-condition classification F1-score of Mistral Large 3, GPT-5.4, and GPT-4o mini on the CT-RATE dataset, evaluated against the dataset’s original labels (extracted with a fine-tuned RadBERT-RoBERTa-4m on 1,000 reports).

B.5 Computation times - oncologic datasets

All the tests for this section were conducted on an Apple MacBook Pro equipped with an Apple M4 Pro chip (12-core CPU: 8 performance cores and 4 efficiency cores) and 24 GB of unified memory. Table 8 presents the computation time for the dataset Colorectal-Liver Metastases, which were consistent with the time on the other datasets.

B.6 Importance of data composition full results

This section provides the complete CAPTOR-100K results referenced in the main text, extending the CAPTOR comparisons of Figures 3–5 and Table 2 to include the larger training-data variant.

B.7 External datasets with original labels

CAPTOR, more specialized for CAP oncological imaging, is significantly outperformed ($p < 0.01$) by our broader 100K model on both CT-RATE [11] and RADchest-CT [33] (Figure 11)

Both our models are outperformed by COL-IPRI [14] and SPECTRE [13] but not significantly for CAPTOR-100K on RADchest-CT.

Country	Studies	Age Median [IQR]	Sex	Contrast	Description
France	4,254	68 [56–77]	F = 53% M = 47%	Injected = 67.8% Non-injected = 13.4% Unknown = 18.8%	Data collected for lesion detection throughout the CAP (incl. adrenal glands, bladder, kidneys, spleen, lungs, lymph nodes, ovaries, rectum, and colon)
Japan	949	70 [35–117]	F = 50% M = 47%	Injected = 66.1% Non-injected = 21.1% Unknown = 12.9%	Data collected for incidental findings of pancreatic cancer.
Spain	4,645	66 [57–75]	F = 47% M = 53%	Injected = 47.8% Non-injected = 10.4% Unknown = 41.8%	Data collected for lesion detection throughout the CAP (incl. the adrenal glands, bladder, kidneys, spleen, lungs, lymph nodes, ovaries, rectum, and colon)
USA	19,852	66 [56–74]	F = 49% M = 51%	Injected = 71.7% Non-injected = 21.9% Unknown = 6.3%	Data collected for disease detection on CAP examinations in an oncologic setting and for lesion detection throughout the CAP (incl. bone, adrenal glands, lungs, gastrointestinal tract, lymph nodes, and pancreas)
Vietnam	193	NA	F = 41% M = 59%	Injected = 80.3% Non-injected = 13.0% Unknown = 6.7%	Data collected for incidental findings of pancreatic cancer.

Table 5: Detailed characteristics of the proprietary dataset by country. IQR = interquartile range, F=female, M=male

Pathology	Whole test set N=10,963	Chest N=4,008	AP N=8,887	CAP N=1,932	Oncological N=3,336	Non- oncological N=7,627
Active bleeding	0.52%	0.04%	0.73%	0.05%	0.09%	0.68%
Aortic aneurysm (unruptured)	5.13%	9.22%	3.13%	7.92%	6.29%	4.72%
Arterial dissection	0.33%	0.19%	0.46%	0.52%	0.39%	0.31%
Arterial/venous occlusion	4.39%	1.99%	5.99%	4.66%	4.71%	4.28%
Common bile duct dilatation	4.48%	-	6.27%	2.48%	2.91%	5.04%
Coronary artery disease / calcification	25.05%	36.65%	-	40.06%	35.82%	21.19%
Epiduritis	0.07%	0.09%	-	0.16%	0.12%	0.05%
Fracture	8.54%	6.86%	10.66%	13.30%	11.81%	7.37%
Hemoperitoneum	1.07%	-	1.51%	0.10%	0.27%	1.35%
Infarction of solid organs	1.59%	0.42%	2.23%	1.09%	0.87%	1.85%
Intestinal obstruction	4.11%	-	5.82%	0.52%	0.99%	5.24%
Intrahepatic bile duct dilatation	6.23%	-	8.73%	2.85%	3.66%	7.15%
Liver fibrosis / Cirrhosis	3.43%	-	4.35%	2.95%	4.92%	2.90%
Mesenteric ischemia	0.60%	-	0.86%	0.21%	0.15%	0.76%
Pancreatitis	1.56%	-	2.22%	0.78%	1.35%	1.63%
Pericardial effusion	5.17%	7.13%	-	7.04%	6.12%	4.83%
Peritoneal carcinomatosis (omental caking)	3.81%	-	5.18%	8.90%	7.97%	2.31%
Peritoneal effusion	16.69%	-	22.99%	8.18%	8.90%	19.48%
Peritonitis	1.94%	-	2.76%	0.21%	0.24%	2.55%
Pleural effusion	14.91%	14.69%	-	18.22%	16.70%	14.27%
Pneumoperitoneum (free air)	2.48%	-	3.47%	0.26%	0.33%	3.25%
Pneumothorax	0.53%	0.30%	-	0.21%	0.15%	0.67%
Pseudo anevrysm	0.23%	0.09%	0.33%	0.26%	0.21%	0.24%
Pulmonary embolism	0.51%	0.55%	-	0.98%	0.99%	0.34%
Pyelocaliceal dilation	4.80%	-	6.34%	4.50%	4.38%	4.96%
Visceral perforation	1.54%	-	2.18%	0.21%	0.18%	2.03%

Table 6: Medical condition proportions in each subset of our whole testing set (internal dataset combined with public datasets)

Dataset	Number of relevant conditions	Merlin	CT-CLIP	SPECTRE	COLIPRI	CTFM
Whole test set	26	+8 / -7	+24 / -0	+9 / -4	+20 / -0	+25 / -0
CAP	26	+11 / -0	+16 / -0	+15 / -0	+13 / -0	+16 / -0
Chest	13	+9 / -1	+8 / -0	+7 / -0	+3 / -0	+8 / -0
AP	20	+3 / -6	+19 / -0	+6 / -6	+15 / -0	+19 / -0
Oncological	26	+13 / -0	+15 / -0	+17 / -0	+15 / -0	+16 / -0
Non-oncological	26	+2 / -9	+23 / -0	+4 / -12	+14 / -1	+25 / -0

Table 7: Number of conditions for which the difference between our model CAPTOR and each baseline is statistically significant. '+' shows the number of conditions where CAPTOR is statistically better, '-' shows the number of conditions where the baseline is statistically better.

Model	Total computation time	Computation time per image
CAPTOR (OURS)	7min38s	2.4s
Merlin	7min49s	2.4s
CT-CLIP	6min34s	2.0s
SPECTRE	1h34min12s	29.1s
COLIPRI	55min28s	17.2
CT-FM	1h09min13s	21.4s

Table 8: Computation times for the dataset Colorectal-Liver Metastases. All computations were conducted on an Apple MacBook Pro equipped with an Apple M4 Pro chip (12-core CPU: 8 performance cores and 4 efficiency cores) and 24 GB of unified memory

Dataset	K	Retrieval text-image					Retrieval image-image					
		CAPTOR (OURS)	CAPTOR-100K (OURS)	Merlin	CT-CLIP	COLIPRI	CAPTOR (OURS)	CAPTOR-100K (OURS)	Merlin	CT-CLIP	COLIPRI	CTFM
Whole test set 10,794	1	307 (266-348)	746 (688-797)	<u>722</u> (622-817)	20.0 (8.88-31.1)	188 (104-281)	1.78 (1.72-1.82)	1.86 (1.79-1.90)	<u>1.82</u> (1.75-1.85)	1.35 (1.27-1.40)	1.63 (1.54-1.68)	1.40 (1.33-1.46)
	5	179 (159-199)	437 (421-452)	<u>353</u> (307-394)	12.2 (6.60-18.7)	98.4 (57.8-144)	1.44 (1.39-1.45)	1.48 (1.43-1.49)	<u>1.46</u> (1.41-1.46)	1.20 (1.12-1.20)	1.35 (1.28-1.36)	1.23 (1.16-1.24)
	10	139 (125-153)	310 (302-319)	<u>235</u> (205-261)	11.3 (6.71-16.6)	70.1 (42.6-101)	1.46 (1.44-1.50)	1.50 (1.47-1.54)	<u>1.48</u> (1.45-1.51)	1.20 (1.16-1.24)	1.35 (1.31-1.40)	1.23 (1.20-1.28)
	50	65.8 (60.9-70.4)	119 (117-121)	<u>77.1</u> (67.8-84.8)	8.02 (4.97-11.8)	30.7 (21.7-40.5)	1.59 (1.57-1.64)	1.63 (1.60-1.68)	<u>1.60</u> (1.58-1.65)	1.24 (1.21-1.31)	1.42 (1.39-1.49)	1.28 (1.26-1.35)
	100	43.1 (40.2-45.8)	72.1 (71.1-73.3)	<u>45.9</u> (40.9-50.2)	6.62 (4.24-9.54)	21.0 (16.0-26.5)	1.61 (1.59-1.66)	1.65 (1.62-1.71)	<u>1.63</u> (1.60-1.68)	1.24 (1.21-1.31)	1.43 (1.40-1.50)	1.29 (1.26-1.36)
Chest 3,944	1	145 (111-197)	219 (175-257)	23.0 (14.0-36.1)	22.0 (10.9-30.4)	<u>184</u> (111-231)	<u>1.33</u> (1.28-1.37)	1.37 (1.30-1.40)	1.27 (1.22-1.31)	1.17 (1.11-1.22)	1.29 (1.22-1.34)	1.18 (1.11-1.23)
	5	85.2 (67.9-113)	134 (118-145)	19.0 (14.2-26.1)	12.8 (7.35-16.7)	<u>92.0</u> (58.4-114)	<u>1.22</u> (1.18-1.25)	1.23 (1.19-1.26)	1.18 (1.14-1.20)	1.11 (1.06-1.14)	1.19 (1.14-1.22)	1.12 (1.07-1.14)
	10	<u>64.3</u> (51.9-84.0)	95.2 (86.2-101)	15.3 (11.9-20.0)	11.6 (7.04-15.0)	63.3 (41.1-78.0)	<u>1.23</u> (1.20-1.27)	1.24 (1.21-1.28)	1.19 (1.16-1.23)	1.11 (1.08-1.16)	1.19 (1.15-1.23)	1.12 (1.08-1.16)
	50	<u>28.6</u> (24.4-35.4)	38.2 (36.2-39.6)	9.10 (7.79-11.1)	8.08 (5.50-9.92)	25.7 (18.8-30.0)	<u>1.26</u> (1.23-1.30)	1.28 (1.24-1.33)	1.23 (1.20-1.28)	1.12 (1.08-1.18)	1.18 (1.14-1.24)	1.13 (1.09-1.18)
	100	<u>18.8</u> (16.2-22.9)	24.4 (23.6-25.0)	6.71 (5.86-7.90)	6.58 (4.65-7.93)	16.8 (13.3-19.1)	<u>1.25</u> (1.21-1.29)	1.26 (1.22-1.31)	1.22 (1.18-1.27)	1.10 (1.06-1.16)	1.16 (1.11-1.22)	1.11 (1.06-1.16)
AP 8,731	1	270 (241-305)	<u>604</u> (560-651)	733 (685-782)	6.00 (2.00-11.0)	29.0 (19.0-39.9)	1.72 (1.66-1.76)	1.80 (1.73-1.82)	<u>1.78</u> (1.71-1.80)	1.25 (1.18-1.29)	1.53 (1.46-1.56)	1.28 (1.23-1.33)
	5	157 (148-168)	<u>357</u> (342-371)	357 (343-372)	4.40 (2.60-6.25)	22.0 (17.8-26.1)	1.41 (1.36-1.42)	1.45 (1.39-1.45)	<u>1.43</u> (1.38-1.44)	1.14 (1.07-1.14)	1.29 (1.23-1.29)	1.17 (1.10-1.17)
	10	122 (116-129)	254 (245-262)	<u>238</u> (230-245)	4.00 (2.80-5.23)	18.2 (15.6-20.7)	1.43 (1.40-1.46)	1.46 (1.44-1.49)	<u>1.45</u> (1.43-1.48)	1.14 (1.11-1.17)	1.30 (1.27-1.33)	1.17 (1.14-1.20)
	50	56.6 (54.8-58.2)	96.6 (94.7-98.4)	<u>78.0</u> (76.2-79.8)	2.92 (2.40-3.42)	12.8 (11.8-13.7)	1.54 (1.53-1.59)	1.57 (1.56-1.62)	<u>1.56</u> (1.54-1.61)	1.16 (1.15-1.21)	1.35 (1.34-1.40)	1.20 (1.19-1.25)
	100	36.9 (36.0-37.8)	58.5 (57.7-59.3)	<u>46.2</u> (45.3-47.1)	2.74 (2.41-3.07)	9.94 (9.36-10.5)	1.56 (1.54-1.61)	1.60 (1.58-1.64)	<u>1.59</u> (1.56-1.63)	1.17 (1.16-1.22)	1.36 (1.34-1.41)	1.20 (1.19-1.25)
CAP 1,881	1	109 (90.3-129)	<u>78.0</u> (61.6-95.7)	25.0 (15.9-34.9)	6.00 (1.99-11.2)	24.0 (14.6-34.3)	1.49 (1.40-1.56)	<u>1.48</u> (1.37-1.53)	1.37 (1.30-1.46)	1.24 (1.15-1.32)	1.36 (1.27-1.43)	1.22 (1.13-1.31)
	5	62.4 (56.1-68.9)	<u>53.6</u> (47.7-60.3)	19.0 (15.3-22.6)	3.60 (2.02-5.44)	13.6 (10.5-17.0)	1.28 (1.22-1.32)	<u>1.27</u> (1.20-1.30)	1.22 (1.16-1.26)	1.13 (1.06-1.16)	1.21 (1.14-1.24)	1.12 (1.05-1.16)
	10	47.8 (44.2-51.6)	<u>38.6</u> (35.3-41.9)	15.1 (12.8-17.4)	3.30 (2.24-4.54)	10.5 (8.55-12.5)	1.30 (1.25-1.35)	<u>1.28</u> (1.23-1.33)	1.24 (1.19-1.29)	1.13 (1.09-1.19)	1.20 (1.16-1.26)	1.13 (1.09-1.19)
	50	19.8 (18.9-20.6)	<u>15.9</u> (15.1-16.8)	8.14 (7.42-8.86)	2.48 (2.04-2.94)	6.98 (6.32-7.62)	1.33 (1.29-1.40)	<u>1.31</u> (1.27-1.37)	1.27 (1.23-1.34)	1.12 (1.08-1.20)	1.19 (1.15-1.26)	1.14 (1.10-1.21)
	100	12.6 (12.2-13.0)	<u>10.8</u> (10.4-11.2)	5.62 (5.20-5.99)	2.14 (1.85-2.44)	5.16 (4.79-5.56)	1.33 (1.28-1.39)	<u>1.31</u> (1.26-1.37)	1.27 (1.23-1.34)	1.12 (1.07-1.19)	1.18 (1.13-1.25)	1.12 (1.08-1.20)
Oncological 3,264	1	224 (195-253)	<u>151</u> (128-174)	38.0 (26.1-51.2)	6.00 (1.99-11.0)	41.0 (29.1-53.9)	1.65 (1.56-1.70)	<u>1.63</u> (1.54-1.68)	1.47 (1.40-1.55)	1.29 (1.20-1.36)	1.37 (1.29-1.44)	1.29 (1.21-1.36)
	5	115 (106-125)	<u>93.8</u> (86.0-102)	27.6 (23.1-32.2)	6.20 (4.19-8.43)	22.8 (18.7-27.1)	1.40 (1.34-1.42)	<u>1.37</u> (1.32-1.41)	1.29 (1.23-1.32)	1.17 (1.10-1.19)	1.23 (1.16-1.25)	1.19 (1.11-1.21)
	10	83.4 (78.4-88.6)	<u>71.7</u> (67.3-76.6)	21.5 (18.8-24.2)	4.60 (3.32-5.90)	16.9 (14.4-19.5)	1.41 (1.37-1.46)	<u>1.39</u> (1.35-1.44)	1.30 (1.26-1.35)	1.17 (1.12-1.22)	1.23 (1.18-1.27)	1.19 (1.14-1.24)
	50	33.5 (32.4-34.6)	<u>28.7</u> (27.6-29.8)	11.2 (10.3-12.0)	3.28 (2.79-3.78)	9.94 (9.16-10.7)	1.48 (1.45-1.55)	<u>1.46</u> (1.43-1.52)	1.38 (1.34-1.44)	1.18 (1.15-1.25)	1.23 (1.20-1.30)	1.21 (1.18-1.27)
	100	21.0 (20.5-21.5)	<u>18.2</u> (17.7-18.7)	8.09 (7.61-8.60)	2.90 (2.56-3.23)	7.38 (6.93-7.88)	1.49 (1.45-1.55)	<u>1.46</u> (1.42-1.52)	1.38 (1.34-1.44)	1.17 (1.13-1.24)	1.21 (1.17-1.28)	1.20 (1.16-1.26)
Non-oncological 7,530	1	190 (160-224)	<u>659</u> (609-711)	745 (619-880)	22.0 (7.78-36.1)	191 (93.7-288)	1.69 (1.61-1.74)	<u>1.78</u> (1.69-1.83)	1.80 (1.72-1.83)	1.27 (1.17-1.35)	1.60 (1.51-1.65)	1.33 (1.23-1.40)
	5	116 (102-132)	377 (361-395)	<u>365</u> (308-423)	14.4 (6.30-22.3)	99.0 (53.6-142)	1.38 (1.31-1.39)	<u>1.43</u> (1.36-1.45)	1.44 (1.38-1.45)	1.15 (1.06-1.17)	1.32 (1.25-1.34)	1.18 (1.10-1.20)
	10	90.3 (80.5-101)	265 (255-276)	<u>236</u> (199-272)	11.9 (5.55-18.2)	69.3 (38.8-98.1)	1.41 (1.37-1.45)	<u>1.47</u> (1.42-1.51)	1.47 (1.44-1.50)	1.16 (1.11-1.21)	1.35 (1.30-1.39)	1.19 (1.14-1.25)
	50	43.0 (39.7-46.7)	95.8 (93.0-98.5)	<u>74.2</u> (64.0-84.8)	8.24 (4.10-12.4)	27.6 (18.7-36.5)	1.52 (1.48-1.58)	<u>1.57</u> (1.53-1.64)	1.58 (1.55-1.63)	1.18 (1.14-1.27)	1.41 (1.36-1.48)	1.23 (1.18-1.31)
	100	28.7 (27.0-30.6)	56.6 (55.3-58.0)	<u>42.8</u> (37.2-48.5)	6.41 (3.27-9.48)	18.7 (13.6-23.5)	1.54 (1.49-1.60)	<u>1.58</u> (1.53-1.66)	1.59 (1.55-1.64)	1.19 (1.14-1.27)	1.42 (1.36-1.49)	1.23 (1.18-1.31)

Table 9: Zero-shot text-image (Recall@K) and image-image (MAP@K) retrieval performance of CAPTOR, CAPTOR-100K and baseline models, expressed as fold-increase over the random baseline (best in bold and second best underlined), by anatomical and oncologic subset. K is the number of retrieved images. SPECTRE [13] could not be evaluated (unavailable projection-layer weights), and CT-FM [17], a vision-only model, was excluded from text-image retrieval.

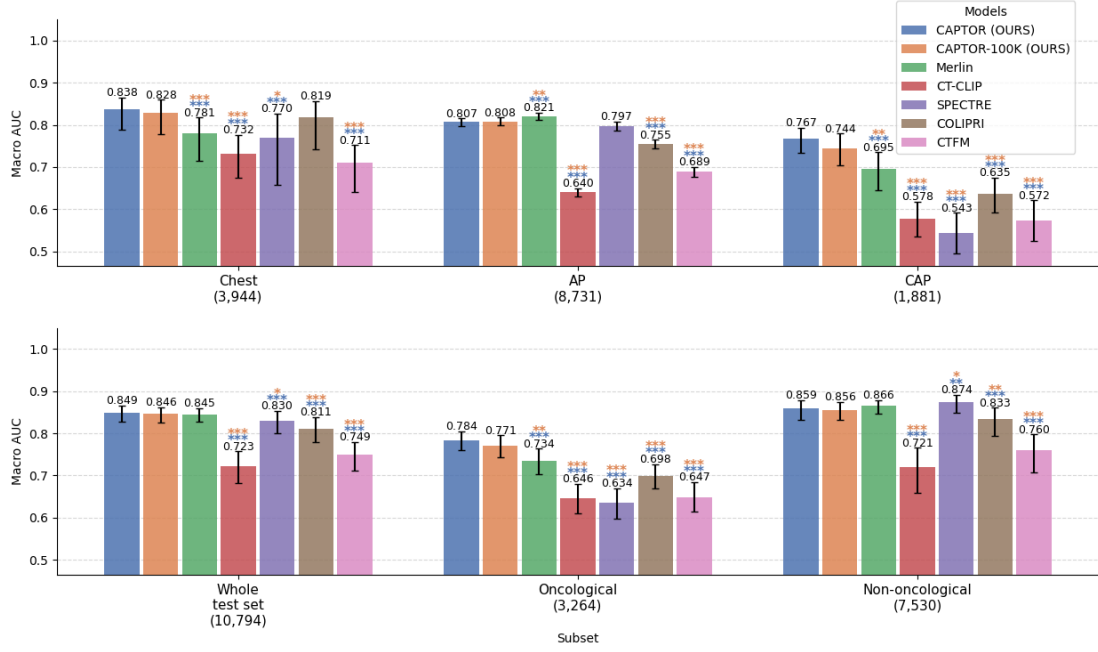


Figure 8: Bar plot showing the classification scores of CAPTOR, CAPTOR-100K and five state-of-the art foundation models, on our test set stratified by anatomical region visible and indication. One XGBoost was trained for each medical condition and model using the image embedding extracted with the model’s image encoder. For each subset and model, the macro-AUC is the unweighted mean of the AUC scores for each medical condition relevant to the subset among the 26 conditions we studied. The confidence intervals and the pairwise comparison were obtained with a patient-level bootstrap and a patient-level permutation test respectively, both of size 2,000. ***= $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, the stars are the color of the comparison model (orange for CAPTOR-100K and blue for CAPTOR)

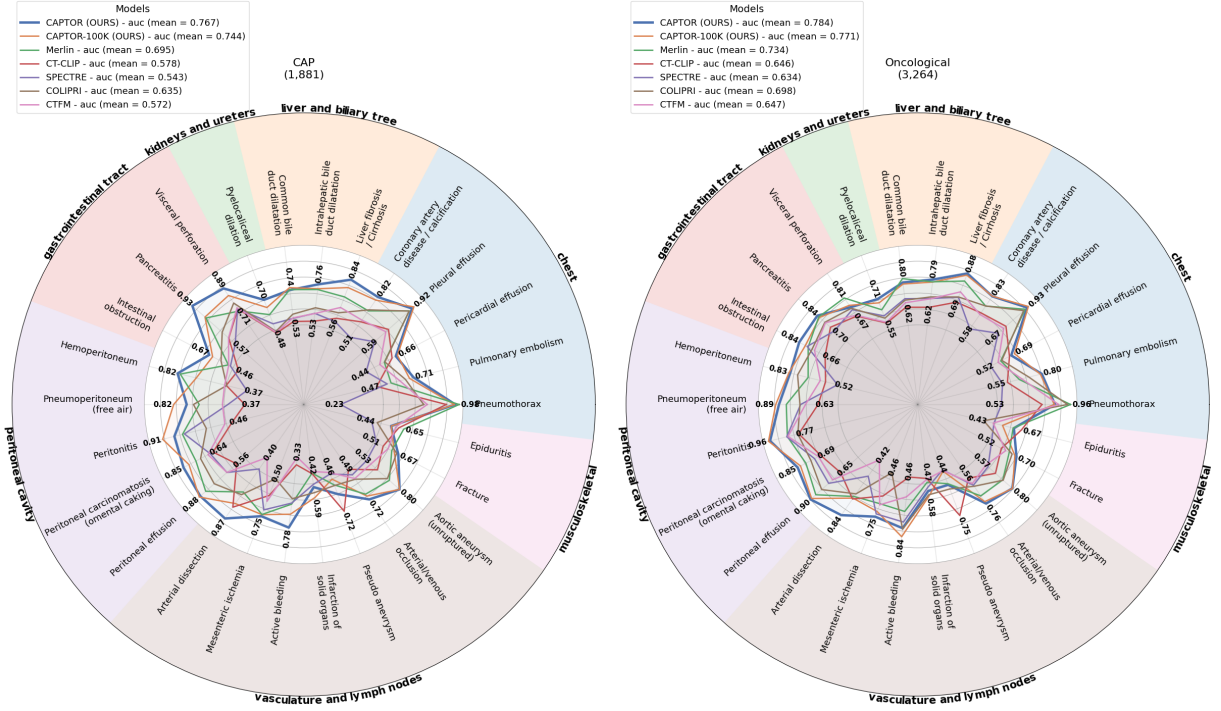


Figure 9: Per-condition classification performance of CAPTOR, CAPTOR-100K, and baselines for the 26 medical conditions, shown separately for the CAP subset (left) and the oncological subset (right). One XGBoost was trained for each medical condition and model using the appropriate image embedding (i.e. the right regional crop for CAPTOR and the right series for the baselines) extracted with the model’s image encoder

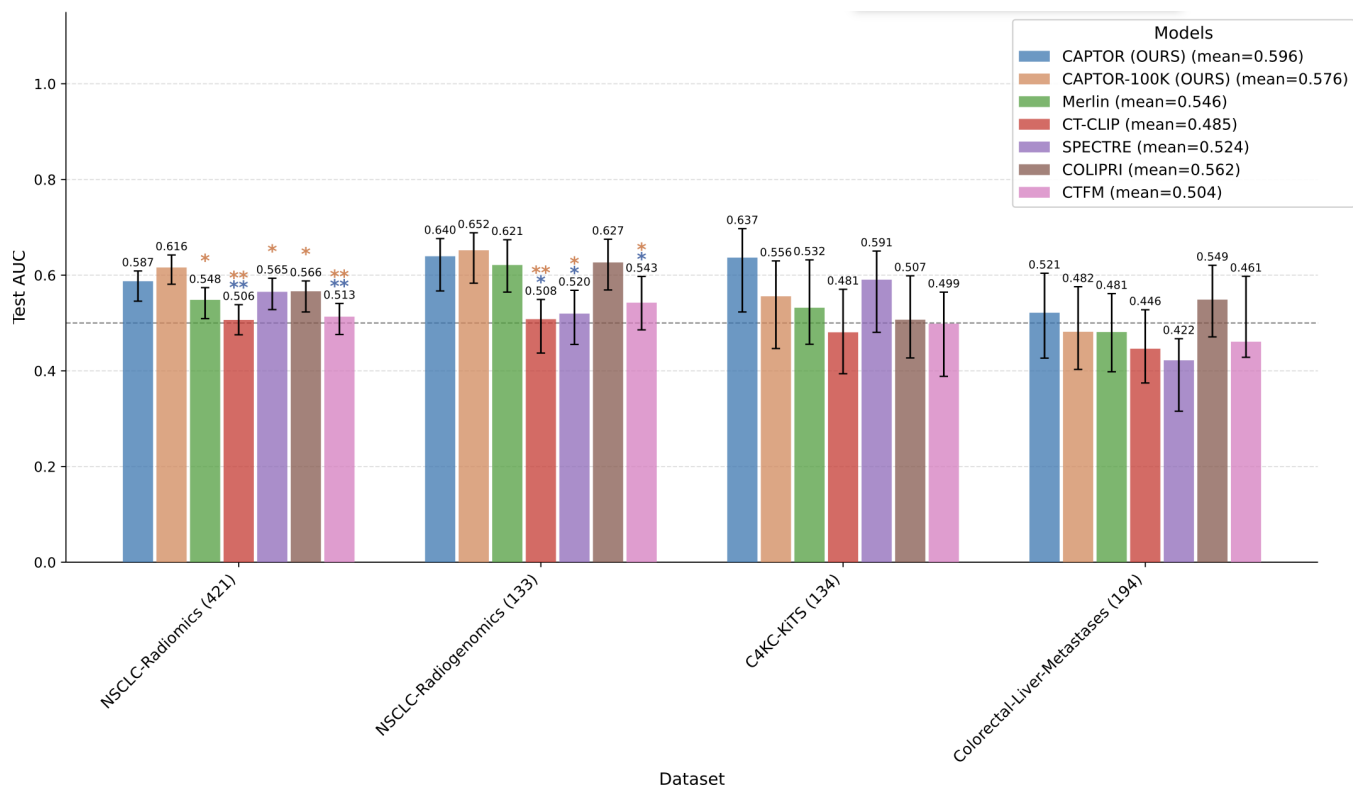


Figure 10: Two-year survival-prediction performance (AUC) of CAPTOR, CAPTOR-100K and five baseline models on four oncologic TumorImagingBench cohorts (NSCLC-Radiomics, NSCLC-Radiogenomics, C4KC-KiTS, Colorectal-Liver-Metastases), using kNN classifiers on frozen image embeddings averaged over 10 random splits. Statistical significance was assessed using a permutation test (2,000 permutations).

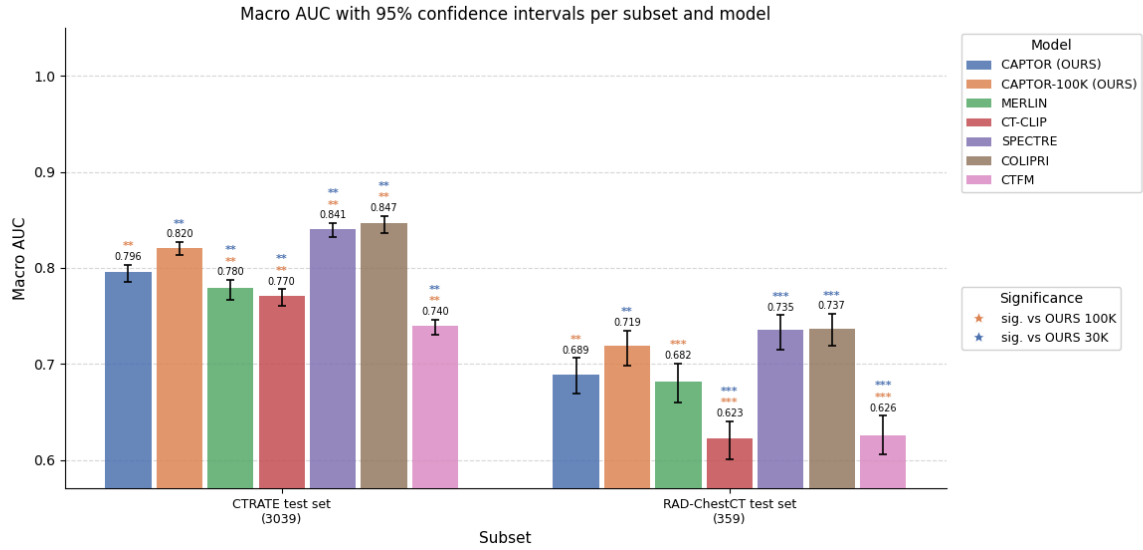


Figure 11: Classification macro-AUC of CAPTOR, CAPTOR-100K, and five baseline models for each abnormality label on the external, non-oncologic CT-RATE [11] and RAD-Chest CT [33] datasets, using each dataset’s original labels. One XGBoost was trained for each medical condition and model using the image embedding extracted with the model’s image encoder. For each subset and model, the macro-AUC is the unweighted mean of the AUC scores for each medical condition relevant to the subset among the 26 conditions we studied. The confidence intervals and the pairwise comparison were obtained with a patient-level bootstrap and a patient-level permutation test respectively, both of size 2,000. *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$