BEYOND THE ENCODER: OBJECT-CENTRIC REPRESENTATIONS EMERGING FROM SEGMENTATION FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models are increasingly used as feature extractors for downstream tasks with limited data, typically through representations obtained from their encoders. However, foundation models trained for structured tasks such as segmentation also form object-specific representations during task execution, whose transferability remains largely unexplored. We investigate Segment Anything Model 2 (SAM2) object pointers, compact object-level representations produced through the segmentation pathway, as transferable representations for downstream tumour prediction. Through experiments on two PET/CT tumour prediction tasks, we find that object pointers learned using segmentation supervision alone already encode tumour-discriminative information, despite never being optimised for the downstream labels. We further introduce task-guided object pointers, where an auxiliary objective jointly shapes the object pointer space for the downstream prediction together with segmentation. Task-guided pointers improve cancer-type classification, outperforming conventional encoder-based transfer and radiomics, while segmentation-only pointers perform best on the more challenging recurrence prediction task. Representation analyses reveal associations between object pointers and interpretable radiomics features, class-dependent organisation of the learned representation space, and explicit spatial grounding around the segmented tumour. Together, these findings suggest that useful transferable representations can emerge through the execution of a foundation model's primary task, providing an alternative to conventional encoder-based transfer.

1 INTRODUCTION

Medical images provide rich quantitative information (Gillies et al., 2016) about tumour phenotype (Fass, 2008) that can support downstream tasks ranging from tumour characterisation to clinical outcomes prediction (Erickson et al., 2017; Bi et al., 2019; Koh et al., 2022; Jiang et al., 2023). Learning representations that capture this information remains challenging, because task-specific imaging datasets are often small and clinical labels are limited. This motivates representations that capture informative tumour characteristics and can transfer across downstream objectives.

Several approaches have been explored to learn informative representations when medical imaging data are limited. Self-supervised learning methods leverage unlabelled data to learn transferable representations (Rani et al., 2024; Shurrab & Duwairi, 2022; Wu et al., 2024), while radiomics represent tumours through a compact set of predefined descriptors of shape, intensity, and texture (Lambin et al., 2012), and have been applied across a range of cancer applications (Aerts et al., 2014; Yip & Aerts, 2016; Avanzo et al., 2020; Liu et al., 2019). More recently, foundation models offer representations learned from large and diverse datasets that can be transferred to data-scarce medical imaging tasks (Moor et al., 2023; Ryu et al., 2025). This has motivated both the adaptation of general-purpose foundation models and the development of domain-specific models for medical image segmentation, classification, and related tasks (Ma et al., 2024; Zhou et al., 2023; Xiang et al., 2025).

However, downstream adaptation of foundation models commonly focuses on representations obtained from the model encoder, either by using it as a frozen feature extractor, or fine-tuning it for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the target task (Huang et al., 2025; Pai et al., 2024; Jacob et al., 2025; Yang et al., 2025). Encoder features are often obtained before much of the computation associated with the model's primary task and therefore precede the task-specific processing that transforms image-level features into more specialised outputs. Representations emerging during this process may provide an alternative source of transferable information that is largely overlooked. This possibility is particularly interesting for segmentation foundation models (Kirillov et al., 2023; Zou et al., 2023; Wang et al., 2023). Segmentation requires the model to identify and delineate a particular object, introducing object-conditioned supervision that transforms generic image features into representation centred on the object of interest. Such representations may therefore encode object-centric characteristics beyond those strictly necessary for delineation. In cancer imaging, this raises a natural question: does the representation formed while segmenting a tumour also encode information useful for predicting its properties?

Segment Anything Model 2 (SAM2) (Ravi et al., 2025), a foundation model for image and video segmentation, provides a natural setting to investigate this hypothesis. In addition to spatial features, SAM2 produces compact *object pointers* through its segmentation pathway to encode segmented objects across frames. Unlike image-encoder features, these representations emerge after prompt- and object-conditioned mask-decoder processing. We therefore investigate object pointers as transferable tumour representations and ask whether information learned through segmentation can support downstream prediction beyond the segmentation task.

We evaluate this hypothesis across two public FDG-PET/CT datasets and distinct downstream prediction tasks: cancer-type classification on FDG-PET-CT-Lesions (Gatidis & Kuestner, 2022) and recurrence prediction on HECKTOR 2022 (Andrearczyk et al., 2023). Our main contributions are:

- We identify SAM2 object pointers as an overlooked source of transferable object-level representations, showing that pointers learned through segmentation alone already encode information predictive of downstream targets.
- We introduce an auxiliary objective for task-guided object pointer learning and show that downstream supervision can further improve these representations, while its benefit depends on the downstream task: task-guided pointers perform best for cancer-type classification, whereas segmentation-only pointers perform best for recurrence prediction.
- We characterise the information encoded by object pointers through radiomics correlation, representation-space organisation, and spatial analysis. We find associations with interpretable tumour characteristics and, for cancer-type classification, information complementary to handcrafted radiomics, alongside class-dependent structure and explicit spatial grounding around the tumour.
- We show that preprocessing choices can substantially alter the downstream information retained by object pointers despite relatively similar segmentation performance, demonstrating that task performance alone does not fully characterise the quality of representations emerging through that task.

2 Related Work

2.1 Foundation Model Representations for Medical Downstream Tasks

Foundation models are commonly adapted to downstream medical imaging tasks through representations obtained from their pretrained encoders, either by using them as frozen features or by fine-tuning the encoder for the target objective (Paschali et al., 2025). This paradigm has been widely studied with general-purpose vision foundation models such as DINO and DINOv2 (Caron et al., 2021; Oquab et al., 2023), whose encoder representations transfer effectively across medical imaging datasets and tasks (Huang et al., 2024; Baharoon et al., 2023; Prokop et al., 2023).

This encoder-centric approach has also been applied to segmentation foundation models. Comparative studies have benchmarked the SAM (Kirillov et al., 2023) image encoder as a feature extractor alongside other foundation models for medical image classification (Huix et al., 2024; Wu & Papiez, 2025), while more targeted approaches have adapted SAM image-encoder representations for downstream tumour classification (Gu et al., 2025; Sun et al., 2024). These studies demonstrate that

representations learned by foundation models can transfer effectively to medical downstream tasks, even when the model was originally developed for segmentation.

However, these approaches largely consider representations obtained directly from the model encoder. For models trained on structured tasks such as segmentation, this motivates investigating representations that emerge beyond the encoder, after the model has performed task-specific processing, as a complementary source of transferable information.

2.2 Segmentation-Guided Representations for Downstream Tasks

Segmentation has also been used to introduce object-specific information into downstream representations. OCCAM (Rubinstein et al., 2025) uses segmentation to isolate objects in pixel space before independently encoding the resulting object-focused images for classification. In medical imaging, SAM-generated masks have been similarly used to construct lesion-focused representations for dermoscopic classification, guiding features from a separate ViT-based classifier (Binzagr & Hariri, 2026). SAM-IE (Wang et al., 2024) further combines SAM-generated masks and contours with the original image to direct classification towards diagnostically relevant regions. These approaches use segmentation outputs to isolate or highlight the object of interest and subsequently inform downstream representation. In contrast, we investigate the object representation formed internally through the segmentation process itself, without constructing a separate representation from the predicted mask.

3 Methodology

3.1 SAM2 Object Pointers

Segment Anything Model 2. SAM2 is a promptable segmentation foundation model for object segmentation in images and videos (Ravi et al., 2025). Given an input image, an image encoder first produces spatial features that are combined with information from previously processed frames through a memory attention. Together with user-provided prompts, these features are processed by the mask decoder to predict segmentation masks for the object of interest. SAM2 then stores information from each processed frame in a memory bank, allowing it to propagate object information across subsequent frames.

Object Pointers. In addition to spatial memory features, SAM2 stores a compact representation of each segmented object, termed as an *object pointer*. Object pointers are derived from the output mask tokens of the mask decoder, which interact with image and prompt features during segmentation. The output token corresponding to the selected mask is projected to obtain the object pointer ($\mathbf{z}_{ptr}$),

$$\mathbf{z}_{ptr} = f_{ptr}(\mathbf{o}_t) \in \mathbb{R}^{256},$$

where $\mathbf{o}_t$ denotes the mask-decoder output token associated with the selected object mask, and f_{ptr} is a learned projection. The resulting object pointer is stored in the memory bank alongside the spatial memory features and can subsequently be attended to when processing the object in later frames.

Unlike image-encoder features, the object pointers emerge after prompt- and object-conditioned segmentation processing and are associated with a specific segmented object. Although SAM2 uses these representations internally to support segmentation across frames, we investigate whether they also encode object characteristics that transfer to downstream tumour prediction.

3.2 SAM2 Representations for Tumour Classification

Problem definition. Given a 3D PET/CT volume $\mathbf{x}$ containing a tumour, our objective is to obtain a tumour representation $\mathbf{z}$ that can be transferred to a downstream classification task. For each subject, the representation is used to predict a class label $y \in \{1, \ldots, C\}$, where C denotes the number of classes for the corresponding downstream task. We formulate the downstream prediction as:

$$\hat{\mathbf{p}}_{cls} = g_\phi\left(f_\theta(\mathbf{x})\right) = g_\phi(\mathbf{z}), \qquad \hat{\mathbf{p}}_{cls} \in [0,1]^C,$$
$$\hat{y}_{cls} = \arg\max\left(\hat{\mathbf{p}}_{cls}\right),$$

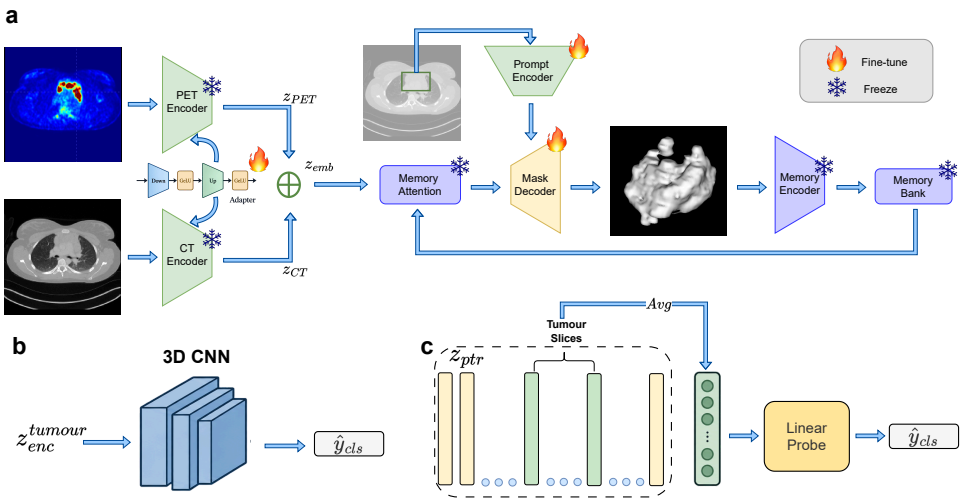

Figure 1: Overview of SAM2 representations for downstream tumour prediction. (a) SAM2 adaptation to 3D PET/CT, using modality-specific encoders and a single bounding-box prompt for tumour segmentation. (b) Encoder-based transfer, where image-encoder features are used by a CNN for downstream classification. (c) Object pointer transfer, where slice-level object pointers produced through segmentation are aggregated into a tumour-level representation and classified using a lightweight linear head.

where f_θ denotes the representation function that maps the input imaging volume $\mathbf{x}$ to a tumour representation $\mathbf{z}$, and g_ϕ is the downstream classification head producing the class probabilities $\hat{\mathbf{p}}_{cls}$. The final predicted class is denoted by $\hat{y}_{cls}$. We compare two sources of representations within SAM2: image-encoder features obtained before object-conditioned segmentation processing, and object pointers emerging through the segmentation pathway. This comparison allows us to assess whether the representations learned during SAM2's primary segmentation task encode transferable information beyond that captured by conventional encoder features.

3D PET/CT segmentation. We adapt SAM2 for 3D clinical PET/CT imaging by treating each volume as an ordered sequence of axial slices, analogous to video frames (see Fig. 1a). Most pretrained SAM2 parameters are frozen, while the prompt encoder and mask decoder are fine-tuned, and lightweight trainable adapters are introduced within the image encoder for parameter-efficient adaptation (Chen et al., 2024). To accommodate paired PET/CT input, we use separate modality-specific image encoders to obtain PET and CT representations, $\mathbf{z}_{PET}$ and $\mathbf{z}_{CT}$, which are fused using a learnable scalar α,

$$\mathbf{z}_{emb} = \alpha \mathbf{z}_{PET} + (1 - \alpha)\mathbf{z}_{CT}.$$

The fused representation is then passed to the SAM2 mask decoder for segmentation. A single perturbed bounding-box prompt derived from the ground-truth annotation is provided on the central axial tumour slice to specify the object of interest, with segmentation subsequently propagated through the volume without additional prompts. Further details on architecture, prompt generation, 3D propagation, and training are provided in Appendix A.1.

Encoder-based representation. As a conventional foundation model transfer baseline, we obtain representations directly from the SAM2 image encoders (see Fig. 1b). Using the ground-truth tumour annotation, we retain the PET and CT encoder features corresponding to tumour-containing axial slices and combine them into a 3D representation, $\mathbf{z}_{enc}^{tumour}$. This representation is provided to a CNN-based classification head,

$$\hat{y}_{cls} = \arg\max(\mathrm{CNN}(\mathbf{z}_{enc}^{tumour})).$$

During downstream training, the image-encoder adapters and classification head are fine-tuned, while the remaining pretrained encoder parameters remain frozen. Further details on the classification architecture and training configuration are provided in Appendix A.2.

Object pointer representation. We instead construct tumour representations from the object pointers produced through segmentation (see Fig. 1c). SAM2 produces an object pointer $\mathbf{z}_{ptr}^{i} \in \mathbb{R}^{256}$ for each processed axial slice i. Using the predicted segmentation, we identify the tumour-containing slices $\mathcal{S}_{tumour}^{pred}$, where $N_{tumour}^{pred} = |\mathcal{S}_{tumour}^{pred}|$ and aggregate them using mean pooling to obtain a single tumour-level representation,

$$\mathbf{z}_{ptr}^{tumour} = \frac{1}{N_{tumour}^{pred}} \sum_{i \in \mathcal{S}_{tumour}^{pred}} \mathbf{z}_{ptr}^{i}.$$

To probe the information already encoded in this representation $\mathbf{z}_{ptr}^{tumour}$, we use a lightweight linear prediction head,

$$\hat{\mathbf{y}}_{cls} = \arg\max(\mathrm{softmax}\big(\mathbf{W}\,\mathrm{Dropout}\big(\mathrm{LN}\big(\mathbf{z}_{ptr}^{tumour}\big)\big) + \mathbf{b}\big)),$$

where LN denotes layer normalisation, and $\mathbf{W}$ and $\mathbf{b}$ are the learnable parameters of the linear layer. SAM2 remains fixed after segmentation training, with downstream labels used only to train the prediction head. Downstream performance therefore probes the tumour-discriminative information already encoded by object pointers learned through segmentation. Further training and optimisation details are provided in Appendix A.3.1.

3.3 Task-Guided Object Pointer Representations

Segmentation-only object pointers are not explicitly optimised for downstream tasks. We therefore investigate whether directly adapting the object-centric representations to a downstream task can improve their predictive information while preserving the segmentation objective. To this end, we introduce an auxiliary classification objective directly on the aggregated tumour-level representation $\mathbf{z}_{ptr}^{tumour}$, using the same lightweight prediction head as in the segmentation-only setting (Sec. 3.2). The segmentation and classification objectives are jointly optimised,

$$\mathcal{L}_{total} = \mathcal{L}_{seg}^{SAM} + \lambda_{\mathrm{cls}}\mathcal{L}_{cls},$$

where $\mathcal{L}_{seg}^{SAM}$ comprises the SAM2 segmentation loss, $\mathcal{L}_{cls}$ the downstream classification loss, and λ_{cls} controls their contribution. Unlike the segmentation-only setting, gradients from $\mathcal{L}_{cls}$ are propagated through the object pointer pathway to the trainable SAM2 components, allowing downstream supervision to directly shape the learned representation while segmentation is jointly maintained.

We therefore consider two complementary settings: segmentation-only object pointers, where downstream labels are used only to train a linear prediction head over a fixed representation, and task-guided object pointers, where downstream supervision additionally shapes the object pointer representation itself. Further training and optimisation details are provided in Appendix A.3.2.

4 Experiments

4.1 Experimental Setup

Datasets and downstream tasks. We evaluate the learned tumour representations on two public FDG-PET/CT datasets with distinct downstream prediction tasks. Both datasets provide paired PET/CT imaging and tumour annotations for segmentation, together with information used to define the downstream tasks.

FDG-PET-CT-Lesions. The FDG-PET-CT-Lesions dataset (Gatidis & Kuestner, 2022) contains whole-body PET/CT imaging with manually annotated tumour lesions. From the full cohort of 900 subjects, we retain 489 subjects with melanoma, lymphoma, or lung cancer, yielding a relatively balanced three-class cancer-type classification task (see Appendix A.4). For subjects with repeated imaging, only the first imaging session is retained, so that each subject occurs once in the cohort. We use this as our primary downstream task.

HECKTOR 2022. The HECKTOR 2022 dataset (Andrearczyk et al., 2023) contains FDG-PET/CT imaging, primary tumour and lymph-node annotations, and clinical outcome information for patients with head and neck cancer collected across seven centres. We define a binary two-year recurrence

prediction task using the primary tumour imaging and available recurrence and follow-up information. Subjects with recurrence within two years are assigned to the positive class; subjects without recurrence and with at least two years of follow-up are assigned to the negative class. Subjects with insufficient follow-up to determine their 2-year recurrence status are excluded. After cohort filtering, 427 subjects remain, including 67 positive cases (see Appendix A.4). This provides a more challenging evaluation setting due to the substantial class imbalance and the dependence of recurrence on factors beyond imaging alone. Although additional clinical variables are available, including smoking history and other patient characteristics, we use only PET/CT imaging to evaluate the information encoded by the learned imaging representations.

Preprocessing and cohort construction. We apply the same preprocessing pipeline to both datasets. To obtain a single tumour representation per subject, we construct unifocal cohorts by retaining the annotated primary tumour for HECKTOR 2022, and the largest connected tumour region for FDG-PET-CT-Lesions. PET intensities are represented as standardised uptake values (SUV), and PET/CT volumes are resampled to a common voxel spacing of $3 \times 2 \times 2$ mm. PET and CT intensities are clipped and globally normalised using statistics estimated from the training data. Axial slices are resized to 128×128 before being passed to the model. Further preprocessing and augmentation details are provided in Appendix A.5.

Training and evaluation protocol. All dataset splits are performed at the subject level and kept fixed across representation approaches. We use five-fold cross-validation, with 80% of the subjects used for model development and 20% held out for testing in each fold. Within the development set, 20% is reserved for validation and used for model selection and hyperparameter tuning.

All SAM2-based approaches are first fine-tuned for tumour segmentation for 20 epochs. The encoder-based and task-guided approaches are initialised from the corresponding segmentation-trained checkpoint (epoch 20), and subsequently trained for the downstream task for up to 20 additional epochs. For the segmentation-only object pointers, SAM2 remains fixed after segmentation training, with the best validation checkpoint used to extract object pointers, while downstream labels are used only to optimise the lightweight prediction head. All downstream classification objectives use class-weighted cross-entropy. Further details on training, model selection, and hyperparameter tuning are provided in Appendix A.7.

Area under the receiver operating characteristic curve (AUC) is the primary evaluation metric, with balanced accuracy (B.Acc.) and F1 score (F1) additionally reported. Segmentation performance is evaluated using Dice similarity coefficient (Dice). Further details on evaluation metrics are provided in Appendix A.8.

Object pointer representations are constructed using predicted segmentation to identify tumour-containing slices, as described in Sec. 3.2. When segmentation fails to identify a tumour, the ground-truth annotation is used only to identify the corresponding slice indices. This ensures that the same subjects are evaluated across approaches, allowing downstream comparisons to focus on representation quality rather than differences in evaluation cohorts caused by segmentation failures.

Baselines. We compare object pointer representations against both learned and handcrafted tumour representations. As a conventional foundation model transfer baseline, we use the encoder-based approach described in Sec. 3.2, where the SAM2 image-encoder features are provided to a CNN prediction head for downstream classification (*Encoder + CNN*). As a handcrafted baseline, we extract radiomics features from the ground-truth tumour masks (*Radiomics*) (Lambin et al., 2012) and provide them to a lightweight linear prediction head analogous to that used for object pointers. Further details on the radiomics baseline are provided in Appendix A.6.

We evaluate both segmentation-only object pointers (*ObjPtr (Seg.)*) and task-guided object pointers (*ObjPtr (Task)*) as described in Sec. 3. To assess whether segmentation-only object pointers contain complementary information to handcrafted tumour characteristics, we additionally concatenate the object pointer and radiomics features into a combined representation (*ObjPtr (Seg.) + Rad.*).

4.2 Experimental Results

Performance in downstream tumour prediction. Table 1 compares object pointer representations with radiomics and conventional encoder-based transfer. On FDG-PET-CT-Lesions, segmentation-only object pointers achieve a mean AUC of 0.781. Although below both radiomics (0.833) and the

Table 1: Downstream prediction performance across representations and datasets. Results are reported as mean ± standard deviation over five-fold cross-validation. The highest mean for each metric is highlighted in **bold**.

Method	FDG-PET-CT-Lesions			HECKTOR 2022		
	AUC↑	B.Acc.↑	F1↑	AUC↑	B.Acc.↑	F1↑
Radiomics	0.833 ± 0.028	0.656 ± 0.046	0.645 ± 0.050	0.518 ± 0.060	0.517 ± 0.027	0.514 ± 0.029
Encoder + CNN	0.836 ± 0.020	0.650 ± 0.060	0.635 ± 0.075	0.504 ± 0.106	0.500 ± 0.000	0.457 ± 0.008
ObjPtr (Seg.)	0.781 ± 0.057	0.611 ± 0.075	0.604 ± 0.072	$\mathbf{0.631 \pm 0.031}$	$\mathbf{0.574 \pm 0.049}$	$\mathbf{0.542 \pm 0.034}$
ObjPtr (Task)	$\mathbf{0.851 \pm 0.017}$	$\mathbf{0.684 \pm 0.024}$	$\mathbf{0.682 \pm 0.023}$	0.584 ± 0.055	0.500 ± 0.000	0.457 ± 0.008
ObjPtr (Seg.) + Rad.	0.849 ± 0.019	0.674 ± 0.033	0.666 ± 0.039	0.549 ± 0.077	0.572 ± 0.066	0.511 ± 0.051

encoder baseline (0.836), this demonstrates that object pointers learned without downstream supervision shaping the representation already encode information predictive of cancer type. Combining segmentation-only object pointers with radiomics increases the AUC to 0.849, exceeding either representation individually and indicating that they contain complementary predictive information for this task. Task-guided supervision further increases performance to a mean AUC of 0.851 with corresponding improvements in balanced accuracy and F1, while largely preserving segmentation performance (Appendix Table 2). The UMAP visualisations similarly show class-dependent organisation in segmentation-only pointers and increased class separation following task-guided training (Appendix Fig. 5a–b).

On HECKTOR 2022, segmentation-only object pointers achieve the highest mean performance across all metrics, including a mean AUC of 0.631 compared with 0.518 for radiomics and 0.504 for the encoder baseline, both close to chance-level discrimination. In contrast to the cancer-type task, combining object pointers with radiomics reduces AUC to 0.549, indicating that the complementary information observed on FDG-PET-CT-Lesions does not extend to recurrence prediction, where radiomics itself provides limited discriminative information. Task-guided object pointers also reduce performance relative to segmentation-only pointers, yielding an AUC of 0.584. For both the encoder and task-guided models, the atypical balanced accuracy and F1 scores reflect collapse towards single-class predictions under the substantial class imbalance. Thus, for this much more challenging and imbalanced outcome-prediction task, downstream supervision does not improve the representation learned through segmentation.

Overall, segmentation-only object pointers encode downstream-discriminative information across both tasks despite receiving no downstream supervision during representation learning. The benefit of task-guided adaptation is task-dependent: it improves cancer-type classification, whereas the segmentation-only representation performs best for recurrence prediction. Across these settings, object pointer approaches outperform conventional encoder-based transfer and handcrafted radiomics representations.

Tumour characteristics captured by object pointers. We next examine whether object pointer representations encode established, interpretable tumour characteristics. We use canonical correlation analysis (CCA) to quantify multivariate associations between the object pointer space and groups of radiomics features comprising tumour shape, first-order intensity, and several texture families (see Fig. 2).

Segmentation-only object pointers (top row) show stronger association with radiomics on FDG-PET-CT-Lesions, particularly for shape- and texture-related feature groups, while considerably weaker associations are observed on HECKTOR 2022. On FDG-PET-CT-Lesions, this alignment indicates that object pointers encode information related to established radiomics tumour characteristics. However, the two representations are not redundant: combining object pointers with radiomics improves cancer-type prediction over either representation alone (see Table 1), providing evidence that they also contain complementary predictive information. On HECKTOR 2022, object pointers exhibit weaker associations with radiomics, whereas radiomics provide limited recurrence discrimination while segmentation-only object pointers remain more predictive. Together, these results show that segmentation-derived object pointers capture information related to established radiomics characteristics while also retaining information beyond these handcrafted features, with the relationship between the two representations varying substantially across downstream settings.

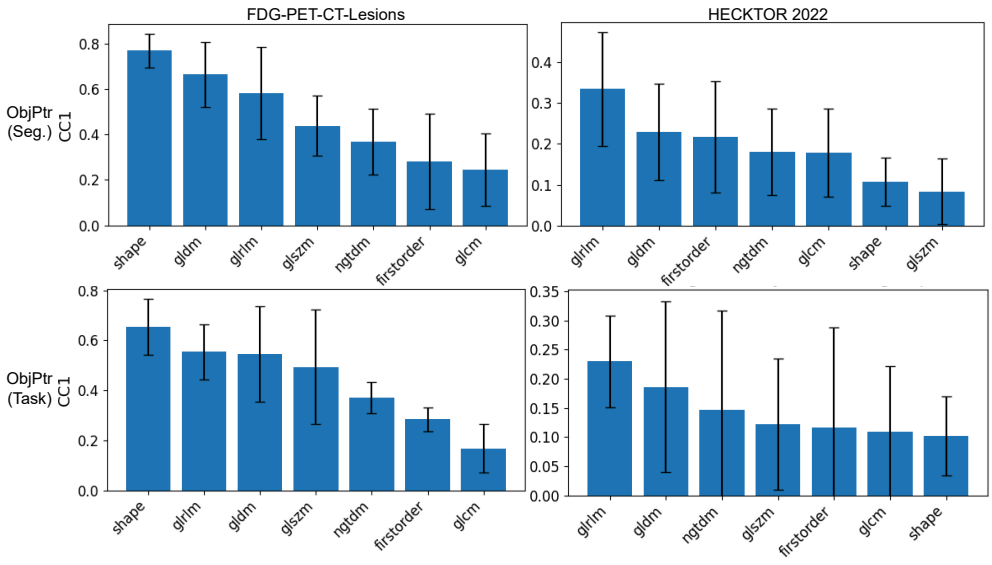

Figure 2: Canonical correlation analysis (CCA) between object pointer representations and radiomics feature groups. Results are shown for segmentation-only and task-guided object pointers across FDG-PET-CT-Lesions and HECKTOR 2022. Bars show the mean canonical correlation over five-fold cross-validation, with error bars indicating standard deviation.

Task-guided training alters these associations, with generally weaker radiomics associations across both datasets (bottom row). Notably, on FDG-PET-CT-Lesions this occurs despite improved downstream classification, indicating that increased task performance does not require stronger associations with the predefined radiomics feature space. On HECKTOR 2022, however, weaker radiomics association is accompanied by reduced downstream performance, suggesting that decreased associations alone should not be interpreted as evidence of a more informative representation. Dimension-wise Pearson correlation analysis further identifies associations between individual object pointer dimensions and specific radiomics features (Appendix Fig. 6).

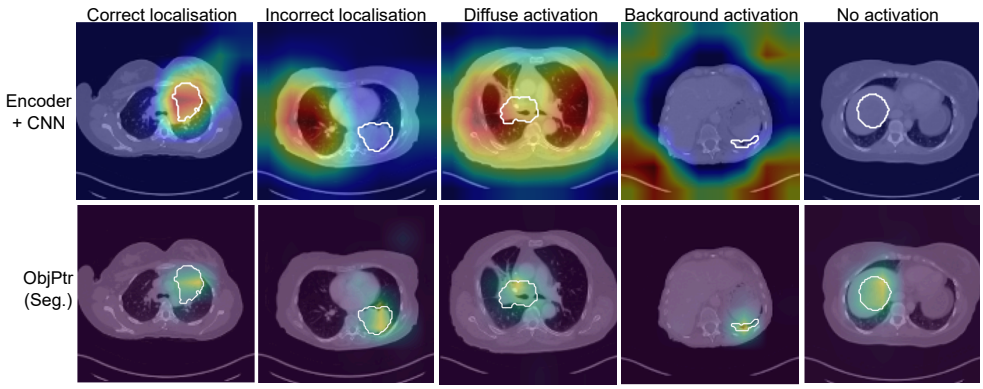

Figure 3: Spatial comparison of encoder-based and object pointer representations. Representative FDG-PET-CT-Lesions cases spanning different spatial activation patterns of the Encoder + CNN baseline, with Grad-CAM maps shown in the top row and corresponding spatial attention maps for segmentation-only object pointers in the bottom row. White contours indicate the ground-truth tumour annotation.

Spatial representation analysis. Unlike image-encoder features, object pointers emerge through object-conditioned segmentation and are therefore explicitly associated with the segmented tumour. Fig. 3 qualitatively illustrates this distinction by comparing Grad-CAM (Selvaraju et al., 2017) from the Encoder + CNN baseline with corresponding spatial attention maps from segmentation-only object pointers across representative cases from FDG-PET-CT-Lesions. While the encoder activations range from well-localised to incorrectly localised, diffuse, or absent responses, the pointer-derived maps remain concentrated around the tumour in these examples. This behaviour reflects the different origins of the two representations: encoder features are obtained before object-conditioned segmentation, whereas object pointers emerge after the model has processed the prompted tumour through the segmentation pathway. Such object conditioning may provide a useful inductive bias for downstream prediction when labelled data are limited.

Effect of PET normalisation on representation quality. We next investigate whether preprocessing choices that affect the segmentation input also influence the information retained by object pointers. We compare global PET normalisation with 3D volume-wise and 2D slice-wise z-score normalisation using segmentation-only object pointers on FDG-PET-CT-Lesions. Downstream object pointer performance decreases progressively from global to volume-wise and slice-wise normalisation, despite comparatively small differences in segmentation performance (see Appendix Table 3). Volume-wise normalisation removes inter-patient intensity variation, while slice-wise normalisation removes variation across slices within each volume, potentially discarding information relevant to downstream tumour discrimination. These results show that preprocessing choices can affect representation quality even when their impact on the primary segmentation task is relatively limited. Consistent with this trend, the UMAP visualisation shows progressively less distinct class-dependent structure from global to volume-wise and slice-wise normalisation (Appendix Fig. 5a,c,d).

Auxiliary loss ablation. We assess the effect of the classification-loss weight λ_{cls} on the task-guided representation on FDG-PET-CT-Lesions. Increasing λ_{cls} does not meaningfully improve downstream classification performance, while segmentation performance progressively decreases (see Appendix Fig. 7). This suggests that placing greater weight on the downstream objective provides no additional predictive benefit while increasingly compromising the segmentation task. We therefore use $\lambda_{\text{cls}} = 1$ in the main experiments.

5 Conclusion

In this work, we investigated SAM2 object pointers as transferable object-level representations for downstream tumour prediction in PET/CT imaging. We show that object pointers emerging from segmentation alone already encode information predictive of downstream targets, despite receiving no downstream supervision during representation learning. The utility of further task-specific adaptation is task-dependent: task-guided object pointers improve cancer-type classification, whereas segmentation-only representations perform best for recurrence prediction.

Our analyses show that object pointers are associated with interpretable radiomics tumour characteristics and exhibit class-dependent organisation within the learned representation space. For cancer-type classification, combining segmentation-only object pointers with radiomics improves performance over either representation alone, indicating complementary predictive information. Qualitative spatial analysis additionally illustrates the tumour-centred nature of representations emerging through object-conditioned segmentation. Finally, we find that preprocessing choices can substantially affect downstream representation quality despite comparatively smaller changes in segmentation performance, highlighting that primary-task performance alone does not fully characterise the representations learned through that task.

Together, these findings demonstrate that useful transferable representations can emerge through the execution of a foundation model's primary task, rather than being confined to its encoder. More broadly, they motivate considering task-emergent internal representations as an additional source of transferable information in structured foundation models. Future work could investigate whether this behaviour generalises across foundation models and imaging domains.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, improving clarity, and refining the presentation of the manuscript. All scientific content, methodological decisions, experiments, analyses, and conclusions were developed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This study uses publicly available, de-identified medical imaging datasets and involves no prospective data collection or direct interaction with patients. We use the data solely for retrospective computational analysis in accordance with the terms of the respective datasets. The proposed methods are intended for research purposes and are not designed or evaluated for direct clinical decision-making. Any clinical translation would require further validation across independent and representative patient populations.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our results. We provide detailed descriptions of the datasets and preprocessing, model implementation and training, and evaluation procedures in the main text and Appendix. Additionally, all datasets used in this paper are publicly available. Anonymous source code is provided in the supplementary material and will be made publicly available upon publication.

REFERENCES

Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):4006, 2014.

Vincent Andrearczyk, Valentin Oreiller, Irene Buvat, Martin Hatt, and Adrien Depeursinge. Overview of the hecktor challenge at miccai 2022: Automatic head and neck tumor segmentation and outcome prediction in pet/ct. In *Head and Neck Tumor Segmentation and Outcome Prediction*, pp. 1–25. Springer, Cham, 2023. doi: 10.1007/978-3-031-27420-6_1.

Michele Avanzo, Lise Wei, Joseph Stancanello, Martin Vallieres, Arvind Rao, Olivier Morin, Sarah A Mattonen, and Issam El Naqa. Machine and deep learning methods for radiomics. *Medical physics*, 47(5):e185–e202, 2020.

Mohammed Baharoon, Waseem Qureshi, Jiahong Ouyang, Yanwu Xu, Abdulrhman Aljouie, and Wei Peng. Evaluating general purpose vision foundation models for medical image analysis: An experimental study of dinov2 on radiology benchmarks. *arXiv preprint arXiv:2312.02366*, 2023.

Wenya Linda Bi, Ahmed Hosny, Matthew B Schabath, Maryellen L Giger, Nicolai J Birkbak, Alireza Mehrtash, Tavis Allison, Omar Arnaout, Christopher Abbosh, Ian F Dunn, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA: a cancer journal for clinicians*, 69(2):127–157, 2019.

Faisal Binzagr and Majed Hariri. Foundation-model-driven skin lesion segmentation and classification using sam-adapters and vision transformers. *Diagnostics*, 16(3):468, 2026.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Tianrun Chen, Ankang Lu, Lanyun Zhu, Chaotao Ding, Chunan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2408.04579*, 2024.

Haoyu Dong, Hanxue Gu, Yaqian Chen, Jichen Yang, Yuwen Chen, and Maciej A Mazurowski. Segment anything model 2: an application to 2d and 3d medical images. *IEEE Transactions on Biomedical Engineering*, 2026.

Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. Machine learning for medical imaging. *radiographics*, 37(2):505–515, 2017.

Leonard Fass. Imaging and cancer: a review. *Molecular oncology*, 2(2):115–152, 2008.

Sergios Gatidis and Thomas Kuestner. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions (fdg-pet-ct-lesions), 2022. URL `https://www.cancerimagingarchive.net/collection/fdg-pet-ct-lesions/`.

Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016.

Pengfei Gu, Haoteng Tang, Islam A Ebeid, Jose A Nunez, Fabian Vazquez, Diego Adame, Marcus Zhan, Huimin Li, Bin Fu, and Danny Z Chen. Adapting a segmentation foundation model for medical image classification. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 167–172. IEEE, 2025.

Shih-Cheng Huang, Malte Jensen, Serena Yeung-Levy, Matthew P Lungren, Hoifung Poon, and Akshay S Chaudhari. A systematic review and implementation guidelines of multimodal foundation models in medical imaging. *Research Square*, pp. rs–3, 2025.

Yuning Huang, Jingchen Zou, Lanxi Meng, Xin Yue, Qing Zhao, Jianqiang Li, Changwei Song, Gabriel Jimenez, Shaowu Li, and Guanghui Fu. Comparative analysis of imagenet pre-trained deep learning models and dinov2 in medical imaging classification. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 297–305. IEEE, 2024.

Joana Palés Huix, Adithya Raju Ganeshan, Johan Fredin Haslum, Magnus Söderberg, Christos Matsoukas, and Kevin Smith. Are natural domain foundation models useful for medical image classification? In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7619–7628. IEEE, 2024.

Athira J Jacob, Indraneel Borgohain, Teodora Chitiboi, Puneet Sharma, Dorin Comaniciu, and Daniel Rueckert. Towards a cardiovascular magnetic resonance foundation model for multi-task cardiac image analysis. *Journal of Cardiovascular Magnetic Resonance*, 27(2):101967, 2025.

Xiaoyan Jiang, Zuojin Hu, Shuihua Wang, and Yudong Zhang. Deep learning for medical image-based cancer diagnosis. *Cancers*, 15(14):3608, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Dow-Mu Koh, Nickolas Papanikolaou, Ulrich Bick, Rowland Illing, Charles E Kahn Jr, Jayshree Kalpathi-Cramer, Celso Matos, Luis Martí-Bonmatí, Anne Miles, Seong Ki Mun, et al. Artificial intelligence and machine learning in cancer imaging. *Communications Medicine*, 2(1):133, 2022.

Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.

Zhenyu Liu, Shuo Wang, Di Dong, Jingwei Wei, Cheng Fang, Xuezhi Zhou, Kai Sun, Longfei Li, Bo Li, Meiyun Wang, et al. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. *Theranostics*, 9(5):1303, 2019.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Suraj Pai, Dennis Bontempi, Ibrahim Hadzic, Vasco Prudente, Mateo Sokač, Tafadzwa L Chaunzwa, Simon Bernatz, Ahmed Hosny, Raymond H Mak, Nicolai J Birkbak, et al. Foundation model for cancer imaging biomarkers. *Nature machine intelligence*, 6(3):354–367, 2024.

Magdalini Paschali, Zhihong Chen, Louis Blankemeier, Maya Varma, Alaa Youssef, Christian Bluethgen, Curtis Langlotz, Sergios Gatidis, and Akshay Chaudhari. Foundation models in radiology: what, how, why, and why not. *Radiology*, 314(2):e240597, 2025.

Jakub Prokop, Javier Montalt Tordera, Joanna Jaworek-Korjakowska, and Sadegh Mohammadi. Deep metric learning for few-shot x-ray image classification. *MedRxiv*, pp. 2023–08, 2023.

Veenu Rani, Munish Kumar, Aastha Gupta, Monika Sachdeva, Ajay Mittal, and Krishan Kumar. Self-supervised learning for medical image analysis: a comprehensive review. *Evolving Systems*, 15(4):1607–1633, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Alexander Rubinstein, Ameya Prabhu, Matthias Bethge, and Seong Joon Oh. Are we done with object-centric learning? *arXiv preprint arXiv:2504.07092*, 2025.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International conference on machine learning*, pp. 29441–29454. PMLR, 2023.

Ji Seung Ryu, Hyunyoung Kang, Yuseong Chu, and Sejung Yang. Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomedical Engineering Letters*, 15(5):809–830, 2025.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022.

Jiakang Sun, Ke Chen, Zhiyi He, Siyuan Ren, Xinyang He, Xu Liu, and Cheng Peng. Medical image analysis using improved sam-med2d: segmentation and classification perspectives. *BMC Medical Imaging*, 24(1):241, 2024.

Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

Changyan Wang, Haobo Chen, Xin Zhou, Meng Wang, and Qi Zhang. Sam-ie: Sam-based image enhancement for facilitating medical image diagnosis with segmentation foundation model. *Expert Systems with Applications*, 249:123795, 2024.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023.

Fuping Wu and Bartlomiej W Papiez. Rethinking foundation models for medical image classification through a benchmark study on medmnist. *arXiv preprint arXiv:2501.14685*, 2025.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22873–22882. IEEE, 2024.

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025.

Zhaochang Yang, Ting Wei, Ying Liang, Xin Yuan, Ruitian Gao, Yujia Xia, Jie Zhou, Yue Zhang, and Zhangsheng Yu. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *Nature Communications*, 16(1):2366, 2025.

Stephen SF Yip and Hugo JWL Aerts. Applications and limitations of radiomics. *Physics in Medicine & Biology*, 61(13):R150–R166, 2016.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023.

A Appendix

A.1 SAM2 PET/CT Segmentation

A.1.1 Architecture and Adaptation

Our model is initialised from the SAM2-Base+ checkpoint, which uses a four-stage hierarchical vision transformer (Hiera) (Ryali et al., 2023) and comprises 80.8M parameters. The PET and CT image encoders are independently initialised from the pretrained SAM2 image encoder, with the original encoder parameters kept frozen throughout training. Modality-specific trainable adapters are inserted at the 12th (middle) and 24th (final) Hiera blocks for parameter-efficient adaptation. Each adapter consists of a down-projection from the original feature dimension to a dimension of 64, followed by a GeLU activation, an up-projection back to the original feature dimension, and a second GeLU activation. The resulting PET and CT embeddings are fused after the image encoder using the learnable weighting parameter α described in the main text, initialised to 0.5. The prompt encoder, mask decoder, image-encoder adapters, and modality-fusion parameter are optimised during segmentation training, while the remaining pretrained SAM2 parameters are kept frozen.

A.1.2 Prompt Generation

SAM2 supports point-, box-, and mask-based prompts. We use a bounding-box prompt as it provides explicit localisation of the tumour while requiring only a simple spatial annotation. A single 2D bounding box is generated from the ground-truth tumour annotation on the central axial tumour slice, which typically contains a large and clearly visible cross-section of the tumour. The bounding box is randomly perturbed by independently expanding each boundary by up to 10% of the corresponding

box dimension. Perturbations are resampled at each training epoch and are also applied during validation and testing, simulating variability in manual box placement and preventing the model from relying on precisely defined prompts.

A.1.3 3D PROPAGATION

Starting from the centrally prompted axial slice, segmentation is propagated through the 3D volume. Propagation is first performed from the central slice towards one end of the volume (backward). The model is then returned to the prompted slice and the memory is reset to prevent information from the backward propagation influencing predictions in the opposite direction (Dong et al., 2026). Propagation is subsequently performed towards the other end of the volume (forward), allowing the tumour to be segmented on both sides of the initial prompt. No additional prompts are provided to subsequent slices.

A.1.4 SAM2 LOSS

We use the standard SAM2 segmentation loss components with adapted relative weighting. The segmentation objective is defined as:

$$L_{seg}^{SAM} = 0.4 \cdot L_{focal} + 0.4 \cdot L_{dice} + 0.1 \cdot L_{IoU} + 0.1 \cdot L_{class},$$

Here, $\mathcal{L}_{focal}$ is a sigmoid focal loss and $\mathcal{L}_{dice}$ is a Dice loss, which together supervise mask prediction. $\mathcal{L}_{IoU}$ supervises the predicted mask-quality score, while $\mathcal{L}_{class}$ supervises object presence. Lower weights are assigned to the IoU and object-presence losses since our setting considers a single object of interest and uses single-mask prediction, reducing the importance of object-presence and mask-selection supervision, respectively.

A.1.5 TRAINING CONFIGURATION

Training largely follows the SAM2 protocol with minor adjustments. We use stochastic gradient descent (SGD) with a momentum of 0.99 and a polynomial learning-rate scheduler with power 0.9. The initial learning rate is set to 0.01 with a weight decay of 3×10^{-5}. Gradients are clipped to a maximum norm of 1.0 for training stability. Models are trained for 20 epochs with a batch size of 8.

A.2 ENCODER-BASED CLASSIFICATION BASELINE

A.2.1 CNN CLASSIFICATION HEAD

The PET and CT image-encoder embeddings are concatenated along the feature dimension, and the embeddings corresponding to tumour-containing axial slices $\mathcal{S}_{tumour}^{GT}$ are retained using the ground-truth tumour annotation,

$$\mathbf{z}_{enc}^{tumour} = \text{Concat}\left(\mathbf{z}_{PET}[\mathcal{S}_{tumour}^{GT}], \mathbf{z}_{CT}[\mathcal{S}_{tumour}^{GT}]\right).$$

The resulting 3D feature volume is passed to a CNN classification head consisting of two $3 \times 3 \times 3$ convolutional layers with 128 and 64 channels, respectively. Each convolutional layer is followed by Group Normalisation, ReLU activation, and dropout ($p = 0.2$). Adaptive average pooling reduces the feature volume to a single $1 \times 1 \times 1$ representation, followed by fully connected layers with 32 and C units, where C is the number of downstream classes. ReLU activation and dropout are applied between the fully connected layers, and a softmax operation produces the final class probabilities. For cases with fewer than three tumour-containing slices, the slice dimension is zero-padded to a minimum depth of three to accommodate the $3 \times 3 \times 3$ convolutional kernels.

A.2.2 TRAINING CONFIGURATION

The encoder-based classification model is initialised from the corresponding segmentation-trained checkpoint. During downstream training, the image-encoder adapters and CNN classification head are optimised, while the remaining pretrained SAM2 parameters remain frozen. The CNN classification head is optimised using AdamW with a learning rate of 1×10^{-3}, weight decay of 1×10^{-2}, and dropout $p = 0.2$. Classification is trained using class-weighted cross-entropy loss with a batch size of 8, and gradients are clipped to a maximum norm of 1.0. The optimal training epoch is selected based on validation performance, as detailed in Sec. A.7.

A.3 OBJECT POINTER CLASSIFICATION

A.3.1 SEGMENTATION-ONLY CLASSIFICATION

For the segmentation-only setting, the object pointer representations obtained from the segmentation-trained model are kept fixed, and only the downstream prediction head is trained. The classifier is optimised using AdamW for up to 100 epochs. Learning rate, weight decay, dropout, and batch size are selected using the validation set, as detailed in Sec. A.7

A.3.2 TASK-GUIDED TRAINING

For task-guided object pointers, the prediction head is trained jointly with the segmentation model. The prediction head is optimised using AdamW with a learning rate of 1×10^{-3}, weight decay of 1×10^{-2}, and dropout $p = 0.2$. The segmentation model is optimised as described in Sec. A.1.5. Training uses a batch size of 8 and gradients are clipped to a maximum norm of 1.0. The optimal training epoch is selected based on validation performance, as detailed in Sec. A.7.

A.4 DATASET AND COHORT DETAILS

A.4.1 FDG-PET-CT-LESIONS

The final FDG-PET-CT-Lesions cohort consists of 489 subjects, including 177 melanoma, 168 lung cancer, and 144 lymphoma cases, resulting in a relatively balanced distribution across the three cancer types.

A.4.2 HECKTOR 2022

Of the 524 subjects in HECKTOR 2022, recurrence information is available for 488. Subjects without recurrence and with less than two years of follow-up are excluded ($n = 48$), as their 2-year recurrence status cannot be determined, leaving 440 subjects. A further 13 subjects without an identified primary tumour are excluded, resulting in a final cohort of 427 subjects. Of these, 67 experience recurrence within two years and 360 have no recurrence with at least two years of follow-up.

A.5 IMAGE PREPROCESSING AND AUGMENTATION

A.5.1 IMAGE PREPROCESSING

PET intensities are represented as standardised uptake values (SUV). PET/CT volumes and corresponding tumour annotations are resampled to a common voxel spacing of $3 \times 2 \times 2$ mm. Volumes are cropped or padded in-plane to 200×200 voxels and limited to a maximum depth of 500 axial slices, while ensuring that the tumour region remains within the retained volume. PET and CT intensities are clipped to the 0.5th–99.5th percentiles and globally z-score normalised using statistics estimated from the training data within each cross-validation fold. Axial slices are resized to 128×128 before being passed to the model, while the out-of-plane dimension was retained to accommodate variable numbers of slices across scans. Following resizing, two very small tumour annotations in FDG-PET-CT-Lesions contained no remaining foreground voxels and were therefore ignored during training and evaluation. No such cases occurred in HECKTOR 2022. As the SAM2 image encoder expects three-channel input, each single-channel modality is replicated across the channel dimension, resulting in $[\mathrm{PET}, \mathrm{PET}, \mathrm{PET}]$ and $[\mathrm{CT}, \mathrm{CT}, \mathrm{CT}]$.

A.5.2 DATA AUGMENTATION

Data augmentation is implemented using MONAI (v1.3.2) (Cardoso et al., 2022) and is primarily spatial to preserve potentially informative PET/CT intensity characteristics. Spatial augmentations are applied jointly to PET, CT, and the tumour annotation to maintain their spatial correspondence. These include random flipping with probability 0.5, random rotation with probability 0.3 and a maximum angle of 30°, and random zooming with probability 0.25 over a range of 0.8–1.2. Gaussian noise with zero mean and standard deviation 0.1 is additionally applied independently to each imaging modality with probability 0.1.

A.6 Radiomics Baseline

A.6.1 Feature Extraction

Radiomics features are extracted from the ground-truth tumour volumes using PyRadiomics (v3.0.1) (Van Griethuysen et al., 2017). PET and CT volumes undergo the same preprocessing described in Sec. A.5 prior to feature extraction. The voxel spacing supplied to PyRadiomics is adjusted to account for the spatial resizing, such that physical voxel dimensions are retained during feature extraction. Features are computed from both PET and CT images and include first-order intensity, shape, GLCM, GLRLM, GLSZM, GLDM, and NGTDM features. For each modality, 107 features were computed and concatenated, with duplicate entries removed, resulting in a 200-dimensional feature vector.

A.6.2 Downstream Classification

The extracted radiomics features are provided to the same lightweight prediction head used for the object pointer representations. The classifier is trained for up to 100 epochs using AdamW, with learning rate, weight decay, dropout, and batch size selected on the validation set as detailed in Sec. A.7.

A.7 Model Selection and Hyperparameter Tuning

All model selection and hyperparameter tuning are performed independently within each cross-validation fold using the corresponding validation set. Held-out test subjects are used only for final evaluation.

A.7.1 Segmentation Model

SAM2 is trained for tumour segmentation for 20 epochs. The checkpoint with the highest validation Dice among the final five epochs is selected for segmentation evaluation.

A.7.2 Encoder + CNN

The encoder-based model is initialised from the final segmentation checkpoint at epoch 20 and trained for up to 20 additional epochs. Here, the segmentation checkpoint serves only as an initialisation before end-to-end downstream fine-tuning, so the final checkpoint from segmentation training is used rather than selecting an intermediate checkpoint based on segmentation performance. The checkpoint achieving the highest validation AUC among the final five epochs is selected for final evaluation.

A.7.3 Segmentation-only object pointers

Object pointer representations are extracted using the segmentation checkpoint with the highest validation Dice among the final five epochs. Since these representations remain fixed during downstream training, the validation-selected segmentation checkpoint is used to provide the best segmentation-derived representation before training the lightweight prediction head. The downstream prediction head is trained for up to 100 epochs. Hyperparameters are selected using a random search over 100 unique configurations. The learning rate is sampled from $\{10^{-4}, 5\times10^{-4}, 10^{-3}, 5\times10^{-3}, 10^{-2}\}$, batch size from $\{8, 16, 32, 64\}$, weight decay from $\{10^{-4}, 5\times10^{-4}, 10^{-3}, 5\times10^{-3}, 10^{-2}\}$, and dropout from $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. The configuration and checkpoint achieving the highest validation AUC are selected independently within each fold.

A.7.4 Task-guided object pointers

The task-guided model is initialised from the final segmentation checkpoint at epoch 20 and trained for up to 20 additional epochs. As in the encoder-based approach, this checkpoint serves only as an initialisation before end-to-end downstream fine-tuning. The checkpoint achieving the highest validation AUC among the final five epochs is selected for final evaluation. The auxiliary classification-loss weight is set to $\lambda_{\text{cls}} = 1$ based on the validation ablation reported in Appendix Fig. 7.

A.7.5 RADIOMICS AND COMBINED REPRESENTATION

The radiomics classifier and the combined radiomics and segmentation-only object pointer classifiers follow the same hyperparameter-tuning and model-selection protocol as the segmentation-only object pointer prediction head, with learning rate, weight decay, dropout, and batch size selected using validation AUC.

A.8 EVALUATION METRICS

For downstream classification, we report area under the receiver operating characteristic curve (AUC), balanced accuracy (B.Acc.), and F1 score. For the three-class FDG-PET-CT-Lesions task, AUC is computed using a one-vs-rest formulation and macro-averaged across classes, while F1 is also macro-averaged. For the binary HECKTOR 2022 recurrence task, the corresponding binary metrics are reported. Segmentation performance is evaluated using the Dice similarity coefficient. All results are reported as mean $\pm$ standard deviation across the five held-out cross-validation folds.

A.9 SEGMENTATION PERFORMANCE

Although segmentation performance is not the primary focus of this study, we report it to assess the segmentation capability underlying the object pointer representations. Appendix Table 2 shows that both segmentation-only and task-guided models maintain good segmentation performance across the two datasets. As expected, Dice is slightly higher for the segmentation-only model, while task-guided training introduces a modest reduction as the representation is additionally optimised for downstream classification. Nevertheless, segmentation performance remains largely preserved, with rare segmentation failures observed across both settings.

Appendix Fig. 4 provides qualitative examples from the segmentation-only model, including representative successful segmentations and failure cases. The failures predominantly occur for very small lesions.

Table 2: Segmentation performance of object pointer models. Dice is reported as mean $\pm$ standard deviation across five-fold cross-validation, alongside the total number of segmentation failures.

Method	**FDG-PET-CT-Lesions**		**HECKTOR 2022**	
	Dice$\uparrow$	**Seg. failures (n)**$\downarrow$	**Dice**$\uparrow$	**Seg. failures (n)**$\downarrow$
ObjPtr (Seg.)	0.737 ± 0.013	9/489	0.693 ± 0.022	3/427
ObjPtr (Task)	0.715 ± 0.010	3/489	0.685 ± 0.017	0/427

A.10 PET NORMALISATION ABLATION

We compare three PET intensity normalisation strategies for the segmentation-only object pointer setting on FDG-PET-CT-Lesions. Global normalisation uses the mean and standard deviation estimated from the training dataset and preserves relative intensity variation between subjects and slices. Volume-wise normalisation independently standardises each 3D PET volume using its own intensity statistics, removing inter-patient differences in intensity scale. Slice-wise normalisation independently standardises each axial PET slice, additionally removing slice-to-slice differences within each volume. Table 3 reports the corresponding segmentation and downstream object pointer performance.

A.11 OBJECT POINTER-SPACE VISUALISATION

To qualitatively examine the organisation of the learned object pointer space on FDG-PET-CT-Lesions, we project the 256-dimensional tumour-level representations into two dimensions using UMAP (McInnes et al., 2018). Each point represents a subject and is coloured according to cancer type. Fig. 5 shows a representative fold, comparing segmentation-only and task-guided object pointers, as well as segmentation-only representations obtained under different PET normalisation

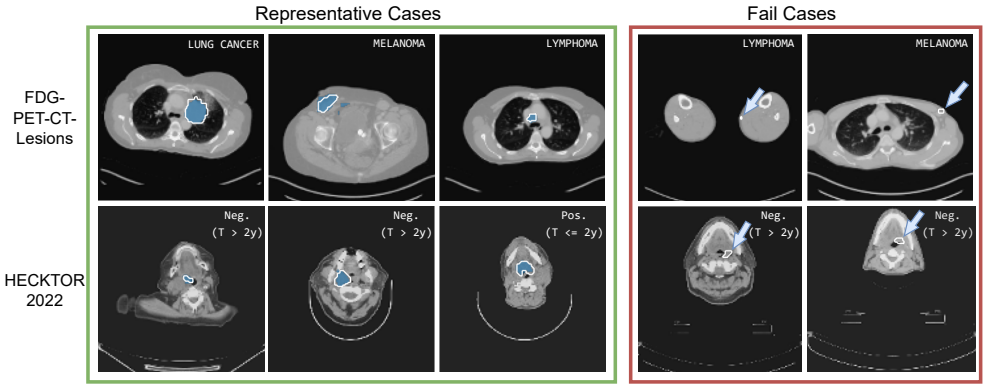

Figure 4: Qualitative segmentation results from the segmentation-only object pointer model. Representative successful segmentations (left) and examples of segmentation failures (right) are shown for FDG-PET-CT-Lesions and HECKTOR 2022. White contours indicate the ground-truth tumour annotation and blue regions indicate the predicted segmentation. Arrows highlight the ground-truth tumour location in failure cases. The corresponding downstream class is shown in the top-right corner of each image.

Table 3: Effect of PET intensity normalisation on segmentation-only object pointer representation quality on FDG-PET-CT-Lesions. Results are reported as mean ± standard deviation across five-fold cross-validation. The highest mean for each metric is highlighted in **bold**.

PET Normalisation	**Classification**			**Segmentation**
	AUC↑	**B.Acc.↑**	**F1↑**	**Dice↑**
Global normalisation	**0.781 ± 0.057**	**0.611 ± 0.075**	**0.604 ± 0.072**	**0.737 ± 0.013**
3D volume-wise normalisation	0.762 ± 0.044	0.567 ± 0.059	0.560 ± 0.056	0.728 ± 0.022
2D slice-wise normalisation	0.723 ± 0.059	0.525 ± 0.067	0.519 ± 0.060	0.703 ± 0.014

strategies. UMAP is fitted independently to each representation using 15 nearest neighbours and cosine distance.

A.12 Radiomics Association Analysis

A.12.1 Canonical correlation analysis

Canonical correlation analysis (CCA) is used to assess multivariate associations between the tumour-level object pointer representation and radiomics features. Radiomics features are grouped according to their respective feature families, including shape, first-order intensity, and texture-based features. Within each cross-validation fold, CCA is fitted using the training subjects and the learned projections are applied to the corresponding held-out test subjects. We report the first canonical correlation, corresponding to the strongest linear association between the two projected feature spaces.

A.12.2 Pearson correlation analysis

To complement the multivariate CCA analysis, we additionally examine associations between individual object pointer dimensions and individual radiomics features using Pearson correlation coefficients (PCC). Correlations are computed between each of the 256 object pointer dimensions and each extracted radiomics feature. Fig. 6 shows the resulting correlation matrices for segmentation-only and task-guided object pointer representations across both datasets.

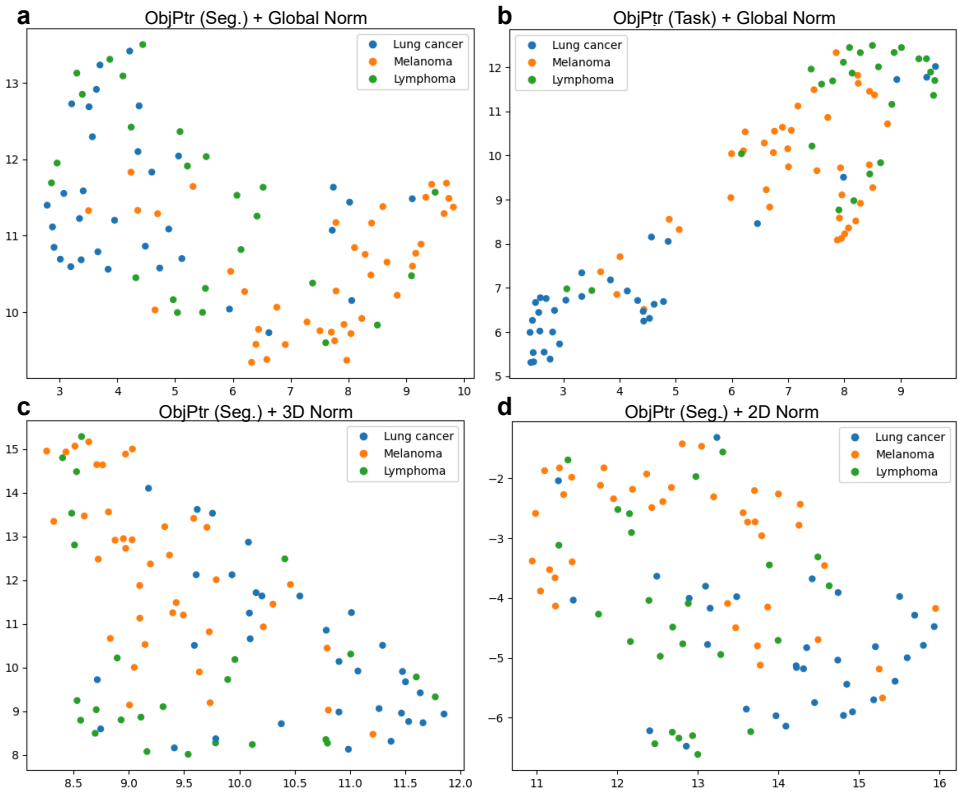

Figure 5: Organisation of the SAM2 object pointer representation space on FDG-PET-CT-Lesions. UMAP projections of tumour-level object pointers for (a) the segmentation-only model with global PET norm, (b) task-guided object pointer training, (c) segmentation-only 3D volume-wise, and (d) segmentation-only 2D slice-wise PET normalisation. Points represent individual tumours and are coloured by cancer type. The segmentation-only representation already exhibits class-dependent organisation, while changes in PET normalisation alter this structure and task-guided training further separates the downstream classes.

A.13 Auxiliary Loss Ablation

We evaluate $\lambda_{\text{cls}} \in \{1, 2, 5\}$ to assess the effect of increasing the contribution of the auxiliary classification objective on FDG-PET-CT-Lesions. Fig. 7 reports downstream AUC and segmentation Dice for each setting on the validation set.

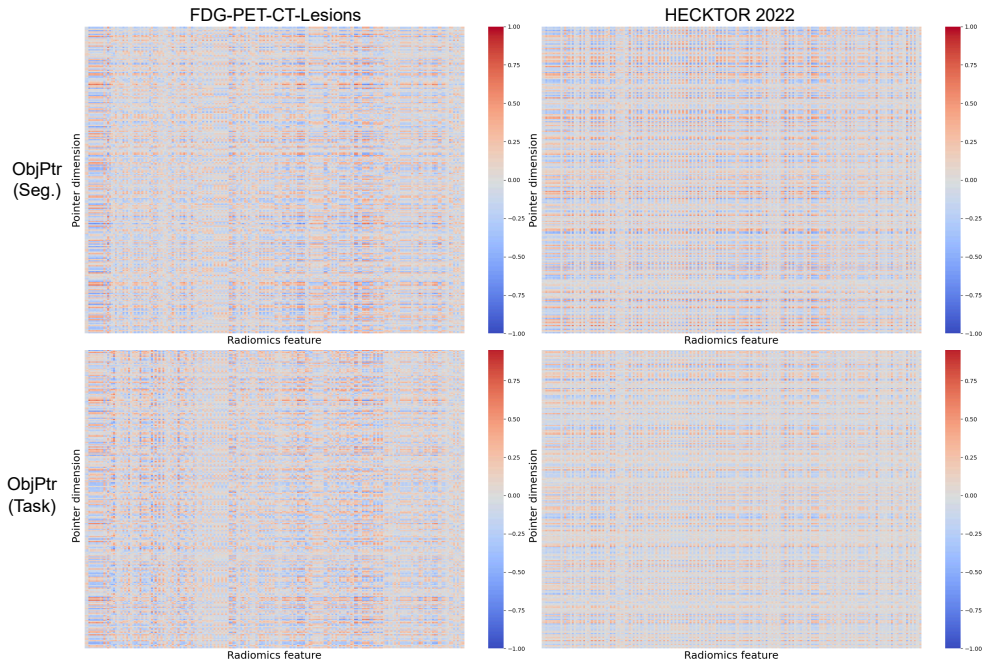

Figure 6: Pearson correlation coefficients (PCC) between individual object pointer dimensions and radiomics features. PCC is shown for segmentation-only and task-guided object pointers across FDG-PET-CT-Lesions and HECKTOR 2022.

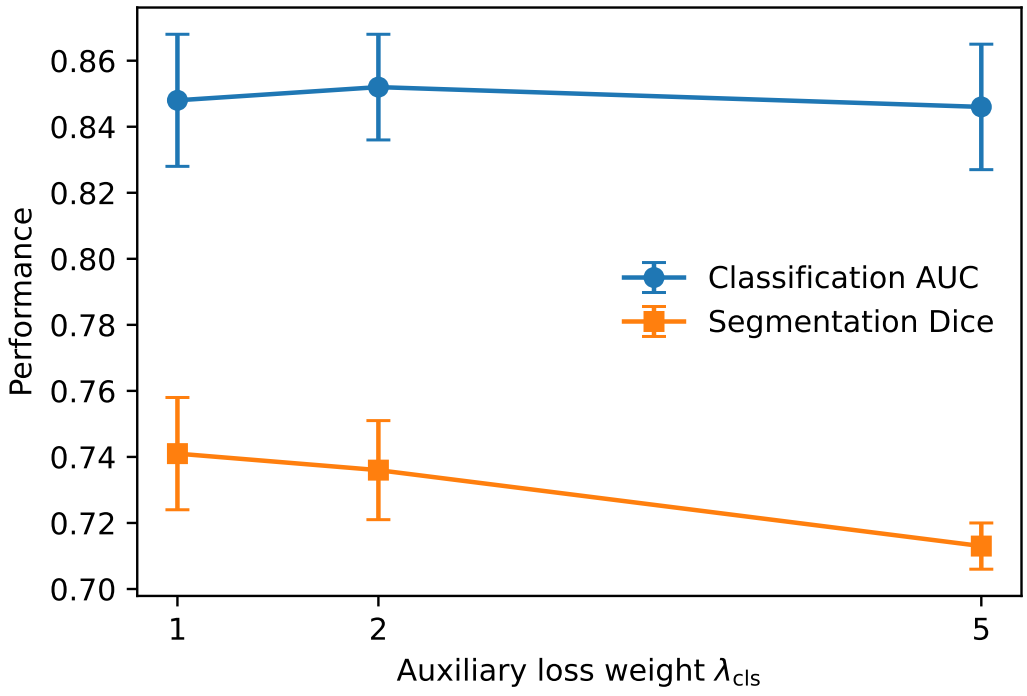

Figure 7: Effect of auxiliary classification-loss weight on FDG-PET-CT-Lesions. Downstream classification AUC and segmentation Dice across different values of λ_{cls}. Points show the mean across five-fold cross-validation, with error bars indicating standard deviation.