npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02087-y

PAM: a propagation-based model for segmenting any 3D objects across multi-modal medical images

Zifan Chen[1,2,7], Xinyu Nan[2,7], Jiazheng Li[3,7], Jie Zhao[4], Haifeng Li[5], Ziling Lin[2], Haoshen Li[2], Heyun Chen[2], Yiting Liu[3], Lei Tang[3] ✉, Li Zhang[2] ✉ & Bin Dong[1,4,6] ✉

Volumetric segmentation is a major challenge in medical imaging, as current methods require extensive annotations and retraining, limiting transferability across objects. We present PAM, a propagation-based framework that generates 3D segmentations from a minimal 2D prompt. PAM integrates a CNN-based UNet for intra-slice features with Transformer attention for inter-slice propagation, capturing structural and semantic continuity to enable robust cross-object generalization. Across 44 diverse datasets, PAM outperformed MedSAM and SegVol, improving average DSC by 19.3%. It maintained stable performance under variations in prompts ($P \geq 0.5985$) and propagation settings ($P \geq 0.6131$), while achieving faster inference ($P < 0.001$) and reducing user interaction time by 63.6%. Gains were strongest for irregular objects, with improvements negatively correlated with object regularity ($r < -0.1249$). By delivering accurate 3D segmentations from minimal input, PAM lowers reliance on manual annotation and task-specific training, providing an efficient and generalizable tool for automated clinical imaging.

Volumetric segmentation is a cornerstone task in medical image analysis[1], involving the precise identification and delineation of regions of interest (RoI) within three-dimensional (3D) medical images. This process is crucial for segmenting various anatomical structures such as organs, lesions, and tissues across a spectrum of imaging modalities, including computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography-computed tomography (PET-CT), and synchrotron radiation X-ray (SRX). The accurate segmentation of these structures is fundamental to a wide array of clinical applications, ranging from disease diagnosis[2–6] and surgical planning[7,8] to monitoring disease progression[9–12] and optimizing therapeutic strategies[13–16]. However, 3D medical images present unique challenges for segmentation tasks, particularly due to the complex inter-slice relationships and the continuous nature of anatomical structures across slices, which are not encountered in traditional 2D image segmentation.

Despite advances in image analysis technology, manual segmentation remains the predominant method in many clinical scenarios[9,10,14,17]. This process is not only time-consuming and labor-intensive but also requires high precision across diverse objects and imaging modalities[18]. These challenges underscore the pressing need for developing semi-automatic or fully automatic segmentation algorithms capable of handling any medical imaging modality and object. Such algorithms have the potential to significantly reduce the time and labor involved while improving the consistency of delineations[19,20].

In response to these challenges, the past decade has witnessed significant advancements in deep learning-based models for medical image segmentation[1,21–24]. These models have demonstrated remarkable capacity to learn complex image features and achieve precise segmentation across various tasks. However, a critical limitation of these approaches is their task-specific nature, often tailored to address particular segmentation challenges posed by specific medical imaging modalities and anatomical structures. This specialization necessitates the creation of large, meticulously annotated datasets for each new task, requiring medical experts to carefully delineate RoIs for specific objects and modalities[25–28]. The process of data collection, manual annotation, and model training must be repeated for each new object or modality[1], an approach that is not only resource-intensive but also impractical for addressing emergent medical scenarios or rare pathologies. The substantial costs associated with data annotation and the scarcity of expert annotations exacerbate these challenges. Consequently, there is a

[1]Center for Machine Learning Research, Peking University, Beijing, China. [2]Center for Data Science, Peking University, Beijing, China. [3]Department of Radiology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing, China. [4]National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing, China. [5]Beijing International Center for Mathematical Research, Peking University, Beijing, China. [6]Beijing International Center for Mathematical Research and the New Cornerstone Science Laboratory, Peking University, Beijing, China. [7]These authors contributed equally: Zifan Chen, Xinyu Nan, Jiazheng Li. ✉e-mail: tanglei@pku.edu.cn; zhangli_pku@pku.edu.cn; dongbin@math.pku.edu.cn

growing demand for more generalized models capable of offering flexibility and rapid adaptability to new tasks without the need for repetitive, extensive training on narrowly defined datasets[29].

The advent of foundation models has revolutionized natural image processing, with the Segment Anything Model (SAM)[30,31] exemplifying remarkable generalization capabilities across diverse tasks. This approach, leveraging user inputs such as points, bounding boxes, and masks, has proven highly effective for natural image segmentation. SAM's success is attributed to its training on vast and varied datasets, which enables it to establish robust correspondences between user prompts and segmentation results in natural images. Inspired by this breakthrough, researchers have begun adapting these versatile frameworks to the medical imaging domain, primarily through two types of models[18,32–44].

Type I models (Fig. 1b), often exemplified by MedSAM[18,32], directly apply the SAM approach to various two-dimensional (2D) medical images. By training on a comprehensive collection of medical images, MedSAM can perform accurate object segmentation within 2D medical images based on simple user-provided 2D prompts. While promising in adapting SAM technology for medical use, Type I models struggle with the complexities of 3D medical imaging. Their lack of consideration for the continuity between adjacent slices in 3D image stacks results in significant challenges in achieving coherent volumetric segmentation. This limitation necessitates more complex user interactions, such as multiple prompts across different anatomical planes or dense annotations on each slice, to achieve satisfactory segmentation results.

Recognizing these gaps, further research has led to the proposal of Type II models (Fig. 1c), such as SegVol[41], which aim to extend the SAM principles to 3D spaces by replacing 2D convolutional kernels with 3D counterparts. This approach enables Type II models to achieve smoother and more accurate 3D segmentation results compared to Type I models. Although Type II models have shown some success in processing volumetric data, they often struggle with generalizing to new, unseen objects or imaging modalities. Moreover, this approach introduces a massive number of parameters, substantially increases computational requirements, and leads to more complex user interactions, posing significant challenges for practical clinical applications.

Upon closer examination of the challenges faced by both Type I and Type II models, we observe that they essentially attempt to directly transplant SAM's approach of modeling correspondences between prompts and segmentations from natural images to 3D medical imaging. While this approach has proven effective for natural images, where objects often have clear boundaries or distinct semantic differences, it faces substantial limitations in medical imaging. Medical images typically exhibit subtle differences in pixel values and textures between objects, making it challenging to achieve generalizability through simple prompt-to-mask alignments. This realization highlights the pressing need to identify and leverage universal characteristics specific to medical images that can boost models' ability to generalize across diverse medical imaging scenarios. Building on this insight, we propose addressing these issues by introducing Type III models (Fig. 1d), which focus on modeling the continuous flow of information in 3D medical images. This concept of continuous flow of information, characterized by inter-slice relationships and the continuous nature of anatomical structures across slices, represents a fundamental distinction between 3D medical images and natural images. By explicitly modeling this characteristic, our approach not only resolves the issue of maintaining continuity when transitioning from 2D prompts to 3D segmentation but also leverages a universal property of medical images, thereby enhancing the model's generalization capabilities.

To implement this approach, we introduce PAM (Propagating Anything Model), an efficient framework to model the continuous flow of information within 3D medical structures. PAM achieves this through two main components: a bounding-box to mask module (Box2Mask), trained on over ten million medical images to respond to bounding-box-style prompts, and a propagation module (PropMask), trained on more than one

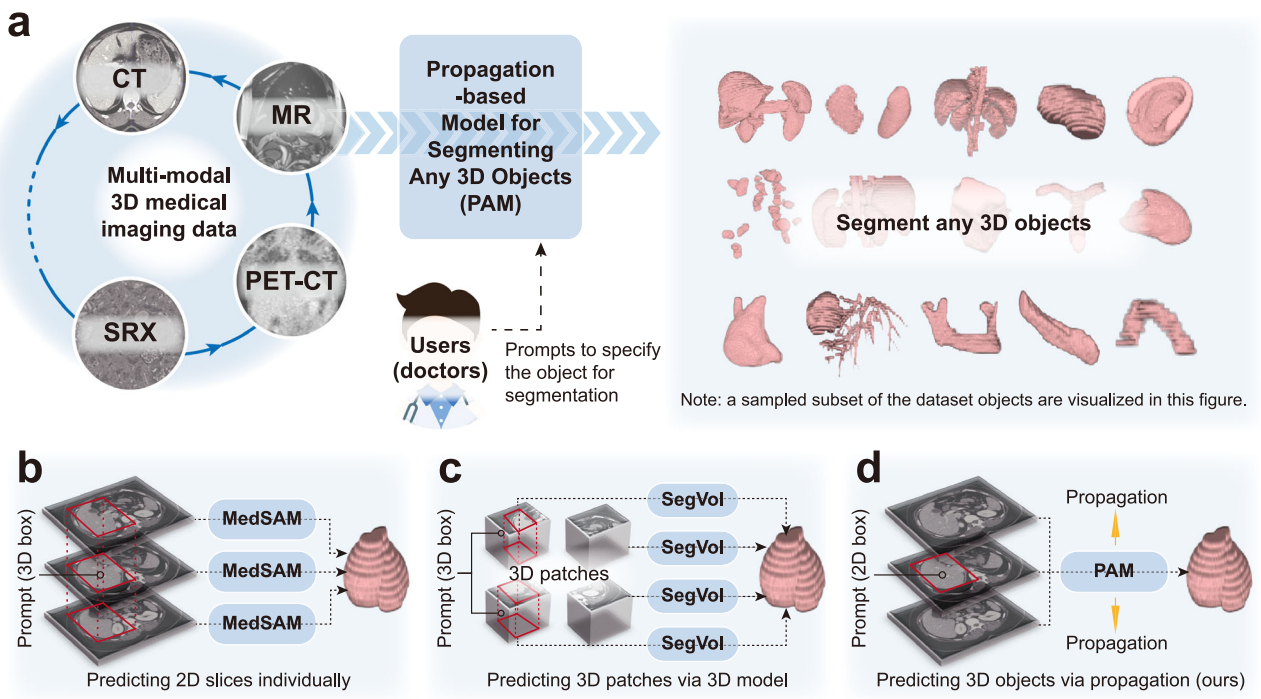

Fig. 1 | PAM is designed for segmenting any 3D objects within various multi-modal 3D medical imaging data. a PAM receives any 3D medical imaging data as input, with users (typically doctors) specifying the target objects for segmentation through prompts. This enables precise and efficient volumetric segmentation of diverse 3D objects, thereby enhancing the efficiency of medical analysis and diagnostics. The Doctor icon was sourced from Iconfont (https://www.iconfont.cn). **b** Type I model: receives a 3D box prompt, predicts each 2D slice using a 2D model, and merges these 2D outcomes into a consolidated 3D prediction. **c** Type II model: receives a 3D box prompt, predicts each 3D patch using a 3D model, and integrates these patch results into a comprehensive 3D prediction. **d** Type III model (ours): receives a 2D box or mask prompt, employs a propagation model to disseminate prompt information throughout the entire 3D space, resulting in a unified 3D prediction.

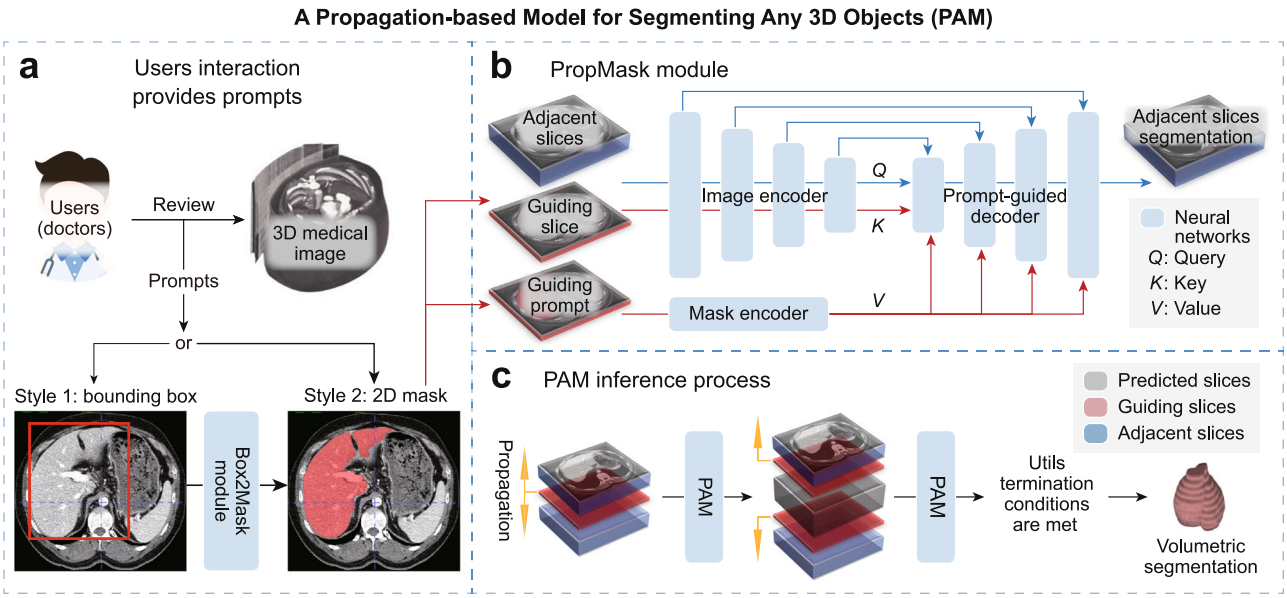

Fig. 2 | Workflow and inference process of the propagation-based model for segmenting any 3D objects (PAM). a User interaction: Users upload a 3D medical image and specify the segmentation target using either a bounding box (style 1) or a 2D mask applied to the largest slice of the target object (style 2). A bounding box is transformed into a 2D mask by the Box2Mask module for standardized processing in the PropMask module. The Doctor icon was sourced from Iconfont (https://www.iconfont.cn). **b** PropMask module: This module conducts volumetric segmentation by propagating information between slices. It begins with the 2D mask and its corresponding image slice (the guiding prompt and slice). Adjacent slices are the targets for segmentation. Image features (*K* and *Q*) are extracted from the guiding and adjacent slices, respectively, using a shared image encoder. The guiding prompt is converted into multi-scale features (*V*) through a mask encoder. These features, along with skip connection features from adjacent slices, are integrated in a prompt-guided decoder to facilitate volumetric segmentation, leveraging the propagation of prompt content across slices. **c** PAM inference: The user provides a guiding slice and prompt. PAM then propagates the prompt information bidirectionally across slices (yellow arrows). This propagation continues until the boundaries of the 3D image are reached or there is no further content to predict, ultimately achieving precise volumetric segmentation. The visualization was generated using ITK-SNAP (version 3.8.0).

million propagation tasks to model inter-slice relationships. The architecture employs convolutional neural networks (CNN) for local segmentation and a Transformer-based attention mechanism for modeling inter-slice information propagation. This hybrid design not only makes PAM more efficient than purely Transformer models (e.g., MedSAM and SegVol) in terms of parameters, computations, and inference speed, but also enables effective information propagation from a single 2D prompt to the entire 3D volume. As PAM's core task is to model structural and semantic relationships between slices, it exhibits generalizability across various medical imaging modalities and to new, unseen objects.

We have rigorously evaluated PAM through comprehensive experiments on 44 medical datasets, covering a variety of segmentation objects and medical imaging modalities. Experimental results demonstrate that PAM consistently outperforms the state-of-the-art (SOTA) segmentation foundation models with greater efficiency. Notably, PAM excels in handling irregular and complex anatomical structures, a common challenge in medical image segmentation. Its ability to capture and propagate intricate structural information allows for accurate delineation of objects with complex shapes, varying sizes, and inconsistent appearances across slices. This capability is particularly valuable in scenarios involving tumors or other anatomical structures with high variability. Moreover, PAM demonstrates robustness to unseen objects and maintains stability across deviated user prompts and different parameter configurations. It also quickly transforms into a powerful expert model for novel object types when fine-tuned with a small amount of annotated data. The fine-tuned PAM notably surpasses proprietary models that are trained from scratch on those limited annotated datasets. These results underscore the potential of PAM as a new paradigm for versatile volumetric medical image segmentation.

PAM: a propagation-based model for volumetric segmentation

PAM focuses on learning the propagation of information across 2D slices in 3D medical images rather than on specific segmentation objects. As depicted in Fig. 2 and Supplementary Fig. S1, the workflow of PAM begins with a user reviewing a 3D medical image and providing prompts within a slice for the target objects. PAM supports two types of prompts: 2D bounding boxes and sketch-based 2D masks (refer to Fig. 2a and Supplementary Fig. S2).

When a 2D bounding box is used, the Box2Mask module executes a foreground segmentation within the box, standardizing the input prompt format as sketch-based 2D masks for subsequent modules. Within PAM, the slice prompted by the user is termed the 'guiding slice,' and the corresponding prompt is known as the 'guiding prompt.' The PropMask module then utilizes the guiding prompt to segment adjacent slices of the same object (Fig. 2b). Initially, the guiding slice and adjacent slices are processed through a shared image encoder to extract image features, forming *K* and *Q*, respectively. Concurrently, the guiding prompt is transformed via a mask encoder into prompt-guided multi-scale features, termed *V*. These features, along with multi-scale features extracted from adjacent slices (as skip connection features), are then integrated in a prompt-guided decoder to produce the final volumetric segmentation. During this process, the PropMask module leverages the continuous flow of information between the guiding slice and adjacent slices to transfer the content of the guiding prompt to the adjacent slices, achieving effective volumetric segmentation.

During inference (Fig. 2c), PAM initially employs the user's prompt for a preliminary round of segmentation. Subsequent rounds utilize the most marginal slices from previous predictions as new guiding slices, enabling the propagation of the segmentation task through adjacent slices. This process continues iteratively until either the boundary of the 3D medical image is reached or when no foreground is predicted in subsequent slices.

Results

Data characteristics and preprocessing

This study utilized 44 publicly available 3D medical image datasets (Supplementary Table S1), encompassing various imaging modalities including CT, MRI, PET-CT, and SRX. These datasets cover 168 different target object types, totaling 1,645,871 3D objects for experimental analysis (Fig. 3a). The

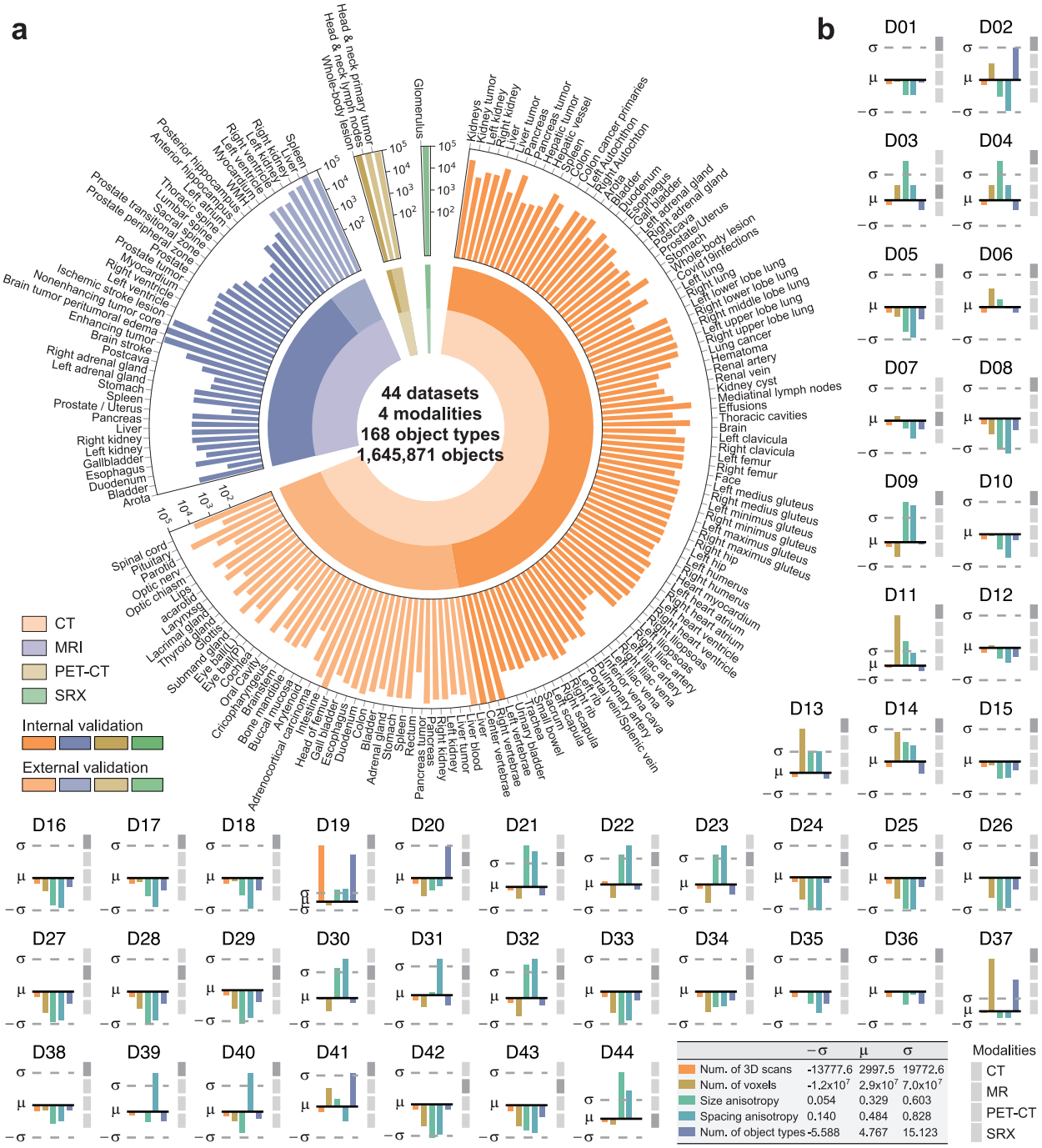

Fig. 3 | Data characteristic across various datasets. a A circular barplot illustrates the range of data modalities and validation splits across multiple datasets. The innermost ring uses distinct colors to represent different medical imaging modalities (orange for CT; blue for MR; yellow for PET-CT; green for SRX). The second ring differentiates between internal and external datasets, with darker shades indicating internal datasets and lighter shades representing external datasets. The outermost layer displays a bar chart that showcases the distribution of segmented object types across the datasets, with quantities log-scaled for optimal visualization. **b** Data fingerprints exhibit the key properties of the 44 datasets used in this study (displayed with z-score normalization over all datasets on a scale of one standard deviation around the mean). See Supplementary Tables S1–S4 for details. The gray bar on the right of each fingerprint encodes the dataset modality, with dark gray positions indicating the active modality corresponding to the legend at the bottom right (e.g., the top block in dark gray denotes CT).

diversity of these datasets is categorized across five dimensions (Supplementary Table S4): number of 3D scans, number of voxels, size anisotropy, spacing anisotropy, and variety of object types. This multidimensional diversity is crucial for a comprehensive evaluation of PAM, as illustrated in Fig. 3b. Size anisotropy, following nnUNet[1], is defined as the ratio of the smallest to the largest size in 3D scans, while spacing anisotropy is calculated as the ratio of the smallest to the largest spacing in 3D scans. In accordance with the protocol established in MedSAM[18], we partitioned these datasets into 34 internal datasets (D01–D34) for training and validation, and ten external datasets (D35–D44) for independent testing (Supplementary Tables S2–S3).

As mentioned in Section 2, PAM comprises two main modules: Box2Mask and PropMask. To train and evaluate the Box2Mask module, a 2D architecture model (detailed in Sections 5.2), we processed 3D images

and their 3D annotations in three steps. First, we simulated bounding boxes based on 3D masks to extract RoI images. Next, we normalized these RoI images. Finally, we applied random data augmentation to enhance the training data. Following this preprocessing, we obtained a total of 19,344,368 samples (2D medical image-mask pairs). These samples were divided into 14,974,620 training samples, 3,782,206 internal validation samples, and 587,542 external validation samples.

To train and evaluate the PropMask module, a 2D architecture model that receives a guiding slice, its prompt, and adjacent slices as input, we further preprocessed 3D images with their 3D annotations. This process involved determining the cropped size to extract both the guiding and adjacent slices, thereby constructing RoI tasks. We then normalized these RoI tasks and employed random data augmentation to enhance their training. Following these preprocessing steps, we amassed a total of 1,345,871 tasks, each consisting of a guiding slice, a guiding prompt, and several adjacent slices. These tasks were distributed as follows: 1,020,576 for training, 258,889 for internal validation, and 66,406 for external validation.

Segmentation performance

We evaluated two versions of PAM: PAM-2DBox, which accepts bounding-box-style (style 1) prompts, and PAM-2DMask, which receives mask-style (style 2) prompts. These were compared against two popular existing models, MedSAM and SegVol, on both internal and external datasets. Unlike PAM-2DBox, which requires only a single-view prompt such as a 2D bounding box, both MedSAM and SegVol necessitate two-view prompts (typically one bounding box on the axial plane and another on the orthogonal plane) within volumetric medical images. These two-view prompts form the tightest possible 3D bounding boxes of the segmentation targets, restricting the inference to the given slice area (Supplementary Fig. S2).

MedSAM, originally designed for 2D medical image segmentation, requires processing each 2D medical slice containing the segmentation targets individually. The results from these segmentations are then stacked to form a volumetric final 3D segmentation, following guidance from its official GitHub (https://github.com/bowang-lab/MedSAM). We refer to this process as 'slice-by-slice prediction.' In contrast, SegVol directly segments volumetric medical images and employs a 'zoom-out-zoom-in' strategy using a resized global image and cropped patches as inputs to balance the acquisition of both global and local image features. We refer to this process as 'patch-by-patch prediction.'

As illustrated in Fig. 4a, our two proposed PAMs exhibit superior segmentation performance, evaluated by the Dice Similarity Coefficient (DSC), across various experimental datasets (Supplementary Table S5). They achieve DSCs of 0.95 or higher on several segmentation objects (e.g., DSC = 0.963 for livers, DSC = 0.950 for kidneys, and DSC = 0.950 for pancreas). Specifically, PAM-2DBox and PAM-2DMask both achieve higher DSCs on 31 of the 34 internal datasets and all ten external datasets compared to MedSAM and SegVol. Overall, PAM-2DBox achieves an average DSC that is 23.1% higher than MedSAM and 19.3% higher than SegVol across all datasets. Similarly, PAM-2DMask demonstrates an average DSC that is 28.5% higher than MedSAM and 24.7% higher than SegVol. These results indicate the robust performance of the proposed PAMs and highlight their outstanding performance on external validation datasets.

Building upon the robust DSC results, a more detailed analysis using additional metrics highlights PAM's superior performance in capturing object boundaries and clinical relevance (Supplementary Tables S7–S11). As shown in Table 1 and Fig. 4b, PAM significantly outperforms MedSAM and SegVol on all three boundary-based metrics (paired t-test $P \leq 0.0078$). For instance, PAM-2DBox achieves a notably lower HD95 (8.937 ± 5.767) and higher NSD (0.709 ± 0.135) compared to MedSAM (HD95: 27.951 ± 39.774; NSD: 0.526 ± 0.263) and SegVol (HD95: 15.457 ± 19.783; NSD: 0.482 ± 0.195). This is particularly important for PAM, as its core design focuses on maintaining structural continuity across slices. Furthermore, our model demonstrates a significant improvement (paired t-test $P \leq 0.0218$) in Sensitivity (PAM-2DBox: 0.738 ± 0.115, PAM-2DMask: 0.845 ± 0.100) compared to MedSAM (0.567 ± 0.240) and SegVol (0.587 ± 0.229). This is crucial from a clinical perspective, where a high sensitivity indicates a lower rate of false negatives, meaning the model is less likely to miss a part of a lesion or other important medical objects. Notably, all models achieve high Specificity values, which is likely because the user prompts effectively focus the models on the regions of interest, thereby effectively reducing false-positive predictions.

The performance gaps highlighted by these metrics stem from the fundamental design principles of the compared models. We observed that MedSAM does not demonstrate superior performance on any 3D segmentation tasks due to its 'slice-by-slice prediction.' SegVol, while showing good performance on organ-related segmentation objects (e.g., DSC = 0.941 for livers, DSC = 0.912 for kidneys, and DSC = 0.842 for pancreas), exhibits a notable decrease in performance on lesion-related or tissue-related segmentation objects (e.g., DSC = 0.189 for whole-body lesions, DSC = 0.001 for white matter hyperintensities, and DSC = 0.161 for glomeruli). These limitations of SegVol stem from the general challenges of 3D segmentation models when trained on limited medical image data. Furthermore, its 'patch-by-patch prediction' strategy can cause fine information loss and discontinuity, posing challenges for predicting lesions with rare annotations and variable shapes. In contrast, both PAM-2DBox and PAM-2DMask can accurately segment organ-related, lesion-related, and tissue-related segmentation objects. For example, their DSCs are 0.669 and 0.755 for whole-body lesions, 0.443 and 0.569 for white matter hyperintensities, and 0.888 and 0.886 for glomeruli, respectively. This indicates the exceptional generalization capabilities of PAM, stemming from its ability to learn generalized tasks (the continuous flow of information between slices rather than specific objects).

Furthermore, to ensure a fairer comparison of the performance of the proposed PAM and the baseline models under the same prompt, we conducted experiments on a representative subset of ten datasets that encompass all modalities and as many medical objects as possible. Given that MedSAM and SegVol require at least a 3D bounding box with two-view prompts, and PAM can also accept these prompts, we compared the performance of each model using the same 3D bounding box prompt (PAM-3DBox). As shown in Fig. 4c and Supplementary Table S17, PAM-3DBox achieves an average DSC of 0.755, demonstrating superior segmentation performance compared to the baseline models (paired t-test $P < 0.001$). Specifically, this represents a 23.9% average absolute DSC improvement over MedSAM and a 28.5% improvement over SegVol. Additionally, compared to our one-view prompt version, PAM-3DBox achieves better overall performance, with a 3.9% increase in average absolute DSC over PAM-2DBox and a comparable performance with PAM-2DMask (paired t-test $P = 0.456$). This is because the 3D bounding box provides more complete information about the target, allowing the model to initiate inference from multiple possible slices, resulting in an effect similar to a voting ensemble.

In addition to prompt-based comparisons, we further included a reference to the fully supervised nnUNet[1], trained independently for each of the ten datasets. As shown in Fig. 4c and Supplementary Table S6, the average DSCs across the ten datasets were 0.716 for PAM-2DBox, 0.767 for PAM-2DMask, and 0.733 for nnUNet. Notably, in dataset D37—which contains only a few samples but a large number of segmentation classes—nnUNet exhibited poor performance due to limited training data. In contrast, PAM, as a general-purpose model, was able to achieve reasonable results despite having never seen the dataset, relying solely on user prompts. To further investigate, we recalculated the average DSCs after excluding D37. On the remaining nine datasets, PAM-2DBox achieved 0.735, PAM-2DMask 0.783, and nnUNet 0.809. These results show that while fully supervised models like nnUNet excel with sufficient labeled data, PAM remains highly competitive, offering strong practical value in realistic scenarios where annotated data are scarce or the segmentation targets are inherently complex.

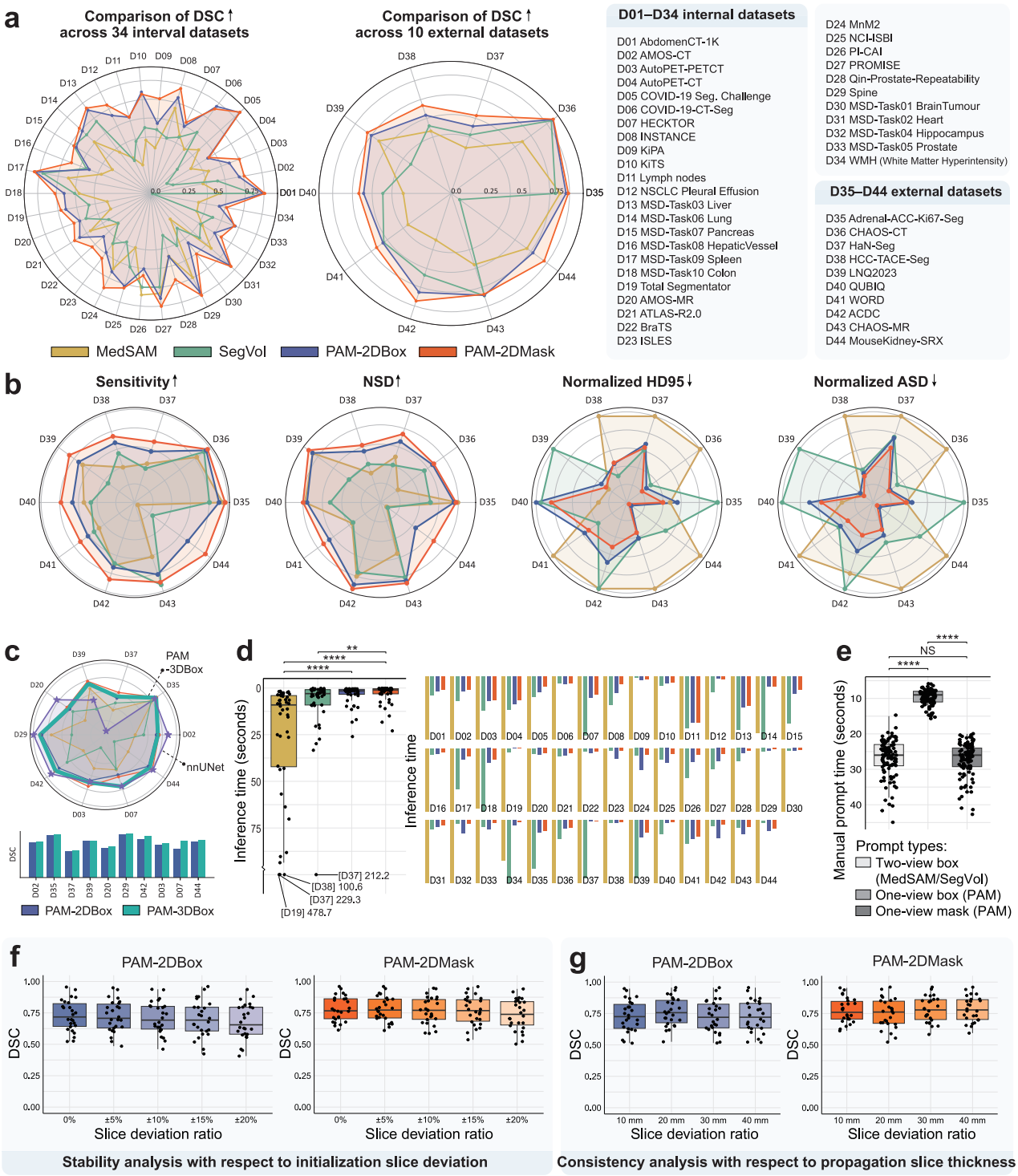

Fig. 4 | Quantitative analysis of PAM across various datasets. a Radar chart comparisons of Dice Similarity Coefficient (DSC) among MedSAM, SegVol, PAM-2DBox, and PAM-2DMask across 44 datasets (D01–D44), with DSC values ranging from 0.0 to 1.0, moving from the center outward. **b** Performance comparison on ten external datasets across four metrics: Sensitivity, Normalized Surface Dice (NSD), 95th percentile Hausdorff Distance (HD95), and Average Surface Distance (ASD). Both HD95 and ASD are normalized by their respective maximum values within each dataset. **c** Comparison of model performance under the same 3D box prompt. The top shows a radar chart comparing PAM-3DBox with baseline prompt-based models using identical 3D box input. The fully-supervised model nnUNet is included as a task-specific performance upper bound. The bottom bar chart illustrates the performance change of PAM-3DBox compared to PAM-2DBox across datasets. **d** Comparison of inference times (seconds). The left side showing a box plot of inference time distribution across 44 datasets, and the right visualizes a comparative analysis of the inference times for each model across these datasets. **e** Comparison of manual prompt times (seconds). A box plot depicts the distribution of interactive prompt times for three distinct prompt types. **f** Stability analysis with respect to initialization slice deviation. Box plots show the DSC distribution for PAM-2DBox (blue) and PAM-2DMask (orange) across deviation levels: 0% (no deviation), ±5%, ±10%, ±15%, and ±20%. **g** Consistency analysis with respect to propagation slice thickness. Box plots show the DSC distribution across thicknesses of 10 mm, 20 mm, 30 mm, and 40 mm. The visualization was generated using ITK-SNAP (version 3.8.0).

Table 1 | Performance comparison of different models on external datasets

Methods	DSC ↑	Sensitivity ↑	Specificity ↑	HD95 ↓	NSD ↑	ADS ↓
MedSAM[18]	0.515 ± 0.156	0.567 ± 0.240	0.994 ± 0.007	27.951 ± 39.774	0.526 ± 0.263	6.383 ± 9.007
SegVol[41]	0.590 ± 0.218	0.587 ± 0.229	0.997 ± 0.004	15.457 ± 19.783	0.482 ± 0.195	4.255 ± 5.377
PAM-2DBox	0.740 ± 0.120	0.738 ± 0.115	0.998 ± 0.002	8.937 ± 5.767	0.709 ± 0.135	1.815 ± 1.058
PAM-2DMask	0.785 ± 0.098	0.845 ± 0.100	0.997 ± 0.004	7.669 ± 5.992	0.774 ± 0.101	1.463 ± 0.897

These quantitative performance analyses underscore PAM's efficacy in accurately segmenting arbitrary 3D objects across a variety of medical imaging modalities and its potential for clinical applications.

Inference and interaction efficiency

We conducted a comprehensive evaluation of the inference times for PAMs, MedSAM, and SegVol across all datasets. As illustrated in Fig. 4d and Supplementary Table S12, MedSAM exhibits the slowest inference speeds (longest inference times), while SegVol shows an improvement over MedSAM. However, our proposed PAMs, in both the 2DBox and 2DMask versions, consistently achieve the fastest inference speeds (shortest inference times) (Wilcoxon rank-sum test, $P < 0.001$; Supplementary Table S13). The superior inference speed of PAMs can be attributed to their unique model structure and efficient inference strategy. PAM employs a hybrid architecture that combines a CNN-based structure, similar to UNet[22], with an attention mechanism inspired by Transformer architectures[45]. This approach allows PAM to leverage the strengths of both architectures for medical image segmentation while maintaining a relatively low parameter count. Specifically, the model has 32.48 M parameters, which increases to 53.1 M parameters when combined with the Box2Mask module for supporting bounding-box prompts. Beyond model design, the inference strategies of the compared methods also differ substantially. MedSAM, a type I model, processes each slice individually using a complex Transformer model (Fig. 1b). SegVol, classified as a type II model, adopts an inference process similar to the 3D-nnUNet model, predicting individual patches with dense overlapping strides that are later merged (Fig. 1c). In contrast, PAM, a type III model, utilizes a 2D model structure and performs bidirectional parallel inference without the need for overlapping window slides (Fig. 1d). These results suggest that PAM's architectural design and inference strategy contribute to its efficiency in segmenting 3D medical images, potentially offering advantages in clinical applications where rapid and accurate segmentation is crucial.

We also explored the interaction efficiency of different models. Both MedSAM and SegVol require two-view prompts, whereas PAM-2DBox only necessitates interaction in one view. In our extracted test subset (see Supplementary Text S1 for experimental details), an experienced radiologist interacted with different datasets and objects using various interaction prompts. We recorded the time taken for each interaction and compared the different prompt types. As demonstrated in Fig. 4e, the one-view box prompt of PAM took significantly less time than the common two-view box prompt used in MedSAM and SegVol (Wilcoxon rank-sum test, $P < 0.0001$; Supplementary Table S14), reducing interaction time by approximately 63.6% and aligning closely with the practical needs of clinical practitioners. Although the one-view mask prompt (used in PAM-2DMask) requires more time than one-view box prompts, it offers an interactive cost comparable to the two-view box prompts ($P = 0.3063$). Furthermore, the more detailed information it provides can greatly enhance the model's overall performance (Fig. 4a–b). These analyses demonstrate the advantages of PAM's two types of prompts over the two-view box prompt, providing users with flexible options for practical application.

Stability and consistency

PAM operates with one-view prompts, typically selected by physicians from the slice with the largest object area, similar to the Response Evaluation Criteria in Solid Tumors (RECIST) guideline. To assess the impact of variations in prompts provided by different physicians on PAM's performance, we conducted a stability analysis. As depicted in Fig. 4f, we simulated deviations from the RECIST-standard confirmed largest slice through five experimental groups: 0% (no deviation), ±5%, ±10%, ±15%, and ±20%. Both PAM-2DBox and PAM-2DMask demonstrated stable DSC across these variations, as confirmed by one-way ANOVA tests, with *P*-values of 0.6736 and 0.5985, respectively (see Supplementary Fig. S9 and Supplementary Table S15 for further details). While performance slightly declined with increasing deviations, it is noteworthy that a deviation of ±20%, which corresponds to a total range of 40%, is uncommon in clinical practice. Even with such substantial deviations, PAMs maintained commendable performance.

Moreover, during the inference process, PAMs iteratively select the most marginal predicted slice as the next round's guiding slice and guiding prompt. The distance of this slice from the original guiding slice could potentially affect the accuracy of subsequent predictions, particularly when far apart, as the propagation relationship weakens with distance. This influence may be amplified through iterative propagation, impacting the overall 3D segmentation of the object. To evaluate the impact of propagation slice thickness, we conducted a consistency analysis with propagation thicknesses of 10 mm, 20 mm, 30 mm, and 40 mm. As shown in Fig. 4g, PAMs maintained predictive stability and consistency across these different thicknesses, as assessed by one-way ANOVA tests with *P*-values of 0.7114 and 0.6131, respectively (refer to Supplementary Fig. S10 and Supplementary Table S16 for more details). Based on these findings, we empirically selected 20 mm as the default propagation thickness for PAMs.

Through these stability and consistency analyses involving varied prompt deviations and propagation thicknesses, PAMs have demonstrated notable predictive stability and consistency. These results provide a reliable foundation for the potential clinical application of PAMs.

Efficacy in segmenting complex and irregular objects

We visualized the qualitative segmentation results of different models as shown in Fig. 5a and Supplementary Fig. S11. Our proposed PAMs effectively utilize propagation information between slices, resulting in visually complete and smooth segmentation outcomes. In contrast, MedSAM, which employs a 'slice-by-slice prediction' and merging strategy, and SegVol, which uses 'patch-by-patch prediction' and integration, do not achieve segmentation visualizations as refined as those produced by PAMs.

We observed that segmentation difficulty varies among different objects. For instance, most organs have relatively fixed shapes, making them easier to learn and segment, whereas some tissues or lesion-related objects present greater challenges. To quantify this process, inspired by BiomedParse[46], we evaluated the 'irregularity' of the objects and the accuracy of our predictions for these irregular objects through three metrics: Box ratio, Convex ratio, and Inverse rotational inertia (IRI), as defined in Section 5.3.4. As demonstrated in Fig. 5b–d, PAM shows performance improvements over MedSAM across most segmentation objects (DSC change > 0). Furthermore, these improvements are more pronounced when the three geometric metrics are smaller ($r < -0.1249$), indicating that PAM is particularly effective at handling irregular objects and more accurately reflects real-world challenges.

To further evaluate our model's clinical utility, particularly for irregular objects such as tumors, we conducted a downstream clinical task as illustrated in Fig. 5e. We utilized pre- and post-treatment CT images from 120 patients with locally advanced gastric cancer. This data was collected from the Department of Radiology at Peking University Cancer Hospital, with the

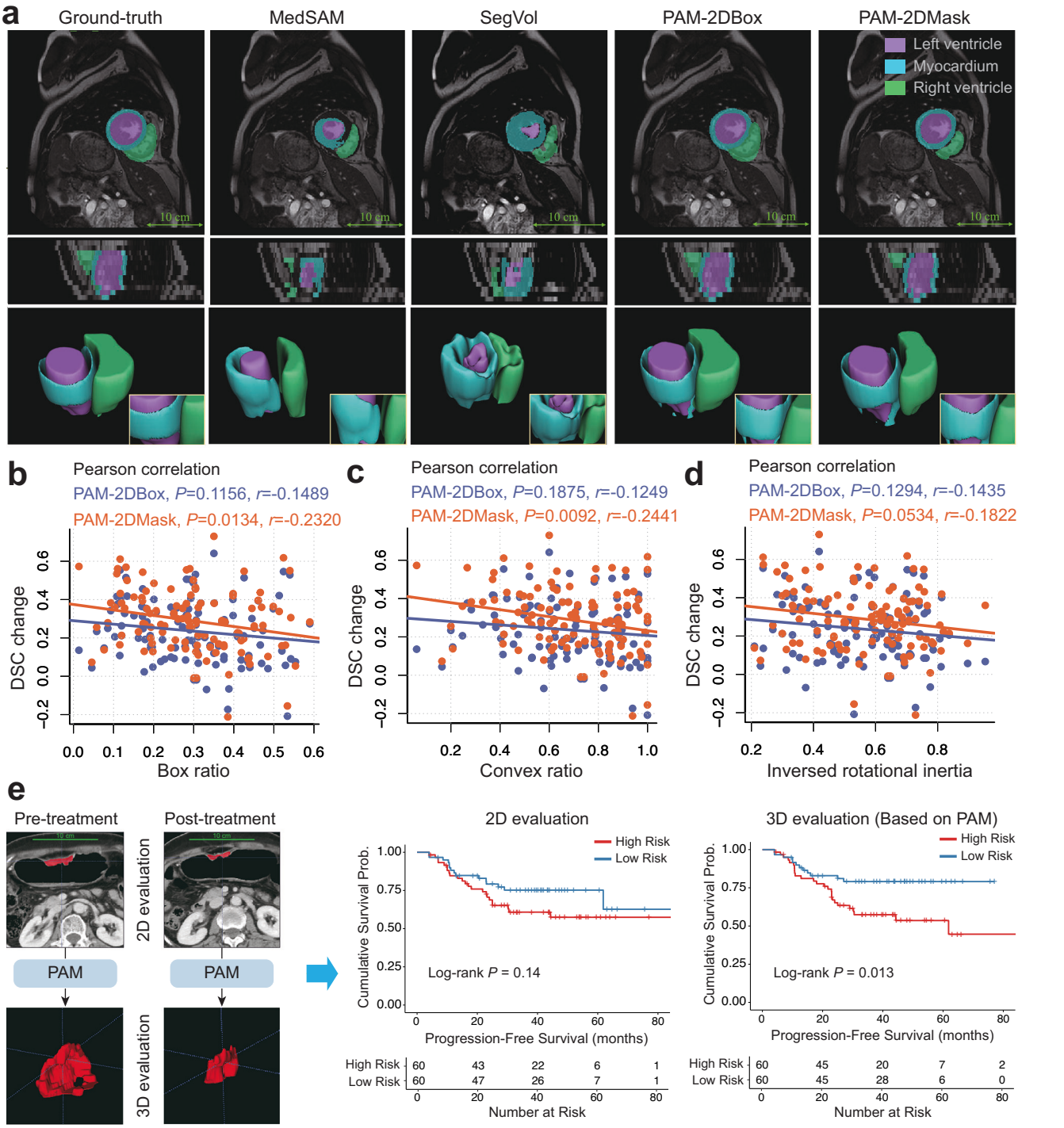

Fig. 5 | Qualitative analysis and the relationship between object shape and performance. a Comparison of segmentation results across various models. From left to right, the columns represent ground truth, MedSAM, SegVol, PAM-2DBox, and PAM-2DMask, respectively. **b–d** DSC change analysis for PAM relative to MedSAM across different geometric metrics. '*r*' refers to the Pearson correlation coefficient, and '*P*' indicates the corresponding *P*-value. **e** Evaluation of PAM's effectiveness in tumor-related downstream tasks on paired pre- and post-treatment CT scans from 120 gastric cancer patients. Physician-provided RECIST-based 2D measurements are used to guide PAM for 3D lesion reconstruction, enabling quantification of area and volume change rates and subsequent Kaplan–Meier survival analysis.

study receiving approval from its Ethics Committee (approval number 2023KT38). In clinical practice, radiologists often use 2D measurements, such as tumor area based on previous studies[47], to assess treatment response and prognosis. Our PAM model, however, offers a more robust approach. Its exceptional generalization and flexibility allow it to use the radiologist's simple 2D measurements as prompts to generate a complete 3D segmentation of the highly irregular tumor region without any fine-tuning. This 3D segmentation enables the calculation of more comprehensive features, such as the ratio of post- to pre-treatment tumor volume. As shown in the Kaplan-Meier survival curves in Fig. 5e, the 3D-based prognostic evaluation using PAM's segmentation can significantly differentiate between high- and low-risk groups (Log-rank $P = 0.013$). In contrast, the traditional 2D-based measurements showed no significant difference (Log-rank $P = 0.14$). This demonstrates PAM's ability to provide more accurate and clinically meaningful insights, which is crucial for patient care and treatment planning.

Beyond this clinical demonstration, these results also highlight the underlying reason for PAM's effectiveness: its robust learning focus, which allows dynamic adaptation to rare and complex shape variations based on inter-slice structural and semantic information. This ability aligns with our initial premise of modeling the continuous flow of information within 3D medical structures. In contrast, MedSAM and SegVol may find these abnormalities challenging due to their reliance on features typical of regular anatomical structures, highlighting the limitations of approaches that do not explicitly account for inter-slice relationships. These analyses demonstrate the potential of PAMs for precise segmentation of various objects, particularly those with irregular shapes.

Generalization and adaptability across diverse segmentation tasks

We explored PAM's generalization capabilities from two perspectives: model fine-tuning and training from scratch. As depicted in Fig. 6a and Supplementary Table S22, PAM, serving as a general segmentation model, outperforms MedSAM on ten external datasets, demonstrating its robust

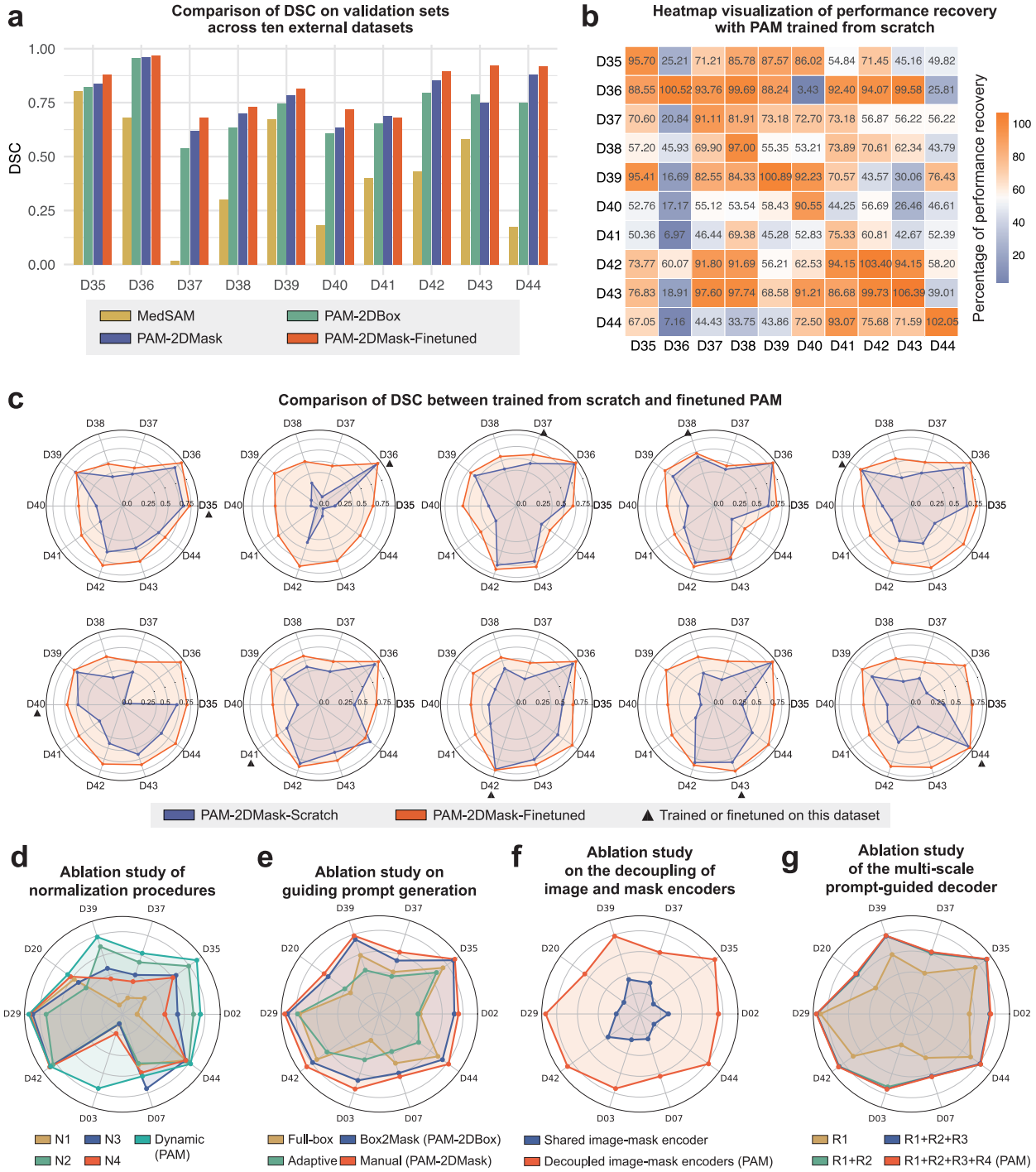

Fig. 6 | Generalization and ablation analysis of PAM. a Bar chart of DSC on ten external datasets comparing MedSAM, PAM-2DBox, PAM-2DMask, and fine-tuned PAM-2DMask. **b** Heatmap of performance recovery when training PAM from scratch, with red indicating higher recovery and blue lower recovery. **c** Radar chart of DSC for scratch-trained versus fine-tuned PAM across D35–D44, with markers denoting datasets used for training or fine-tuning. **d**–**g** Ablation studies of PAM on a representative subset consisting of ten datasets, including normalization strategies, guiding prompt generation, encoder decoupling, and multi-scale prompt-guided decoding.

generalization ability to unfamiliar datasets and objects. We then partitioned these ten datasets into training and validation sets and conducted minimal fine-tuning of PAM on the training sets to create the PAM-2DMask-Finetuned model. Experiments indicate that with minimal data fine-tuning, PAMs can quickly adapt to corresponding tasks and enhance performance. However, the improvement of PAM-2DMask-Finetuned over PAM-2DMask is less significant compared to PAM's improvement over MedSAM (paired t-test, $P = 0.0011$), suggesting that the general model of PAM already performs well on unseen objects.

We also trained PAM from scratch on the divided training set and evaluated the performance recovery percentage relative to the general model PAM across ten datasets (as shown in Fig. 6b). The experiments reveal that even with limited data, training PAM from scratch can achieve over 75.33% performance recovery on corresponding datasets, indicating that PAM's learning tasks are sufficiently straightforward to allow rapid adaptation on limited samples. Furthermore, we observed that segmentation objects with similar structures to the training objects benefited in performance. For instance, when trained from scratch on dataset D35, the performance recovery on D39 reached 87.57% due to the objects in these two datasets being both lesion-related and structurally similar. This further underscores that PAM's learning focus is not on specific semantic objects, but rather on the structural or semantic information transfer relationships between slices, aligning with our aim to model the continuous flow of information within 3D medical structures.

Using radar charts (Fig. 6c), we compared the segmentation performance on ten datasets of the fine-tuned model (PAM-2DMask-Finetuned) and a model trained from scratch (PAM-2DMask-Scratch). As noted, aside from performing well on the datasets where it was fine-tuned, PAM-2DMask-Scratch also achieved commendable performance on datasets with similar structural segmentation objects. Generally, the performance of PAM-2DMask-Finetuned is superior to PAM-2DMask-Scratch, indicating that the general capabilities aid in fine-tuning specific objects and achieving more precise segmentation results[48,49]. This also demonstrates PAM's general capability to handle various segmentation objects, reinforcing its effectiveness in capturing inter-slice relationships across diverse medical imaging scenarios.

These analyses highlight PAM's robust generalization capabilities, showcasing its effectiveness in both fine-tuned and from-scratch training scenarios across diverse datasets and structural variations. The results affirm PAM's ability to model the continuous flow of information within 3D medical structures, enabling adaptability and precision in segmentation tasks across various imaging modalities and object types.

Ablation studies

To better understand the contribution of different components in PAM, we conducted a series of ablation studies, focusing on the normalization procedure and the architectural design.

As medical images from different anatomical regions exhibit diverse intensity distributions, normalization is critical for robust segmentation. We compared PAM's dynamic normalization (refer to Section 5.3.1 for details) with four fixed normalization strategies commonly used in clinical practice (Supplementary Table S18). As shown in Fig. 6d and Supplementary Table S19, fixed normalization works for certain anatomical structures but cannot accommodate the diversity of objects encountered in general-purpose segmentation. In contrast, the dynamic normalization used in PAM consistently adapts to each evaluation instance and achieves superior performance across eight of the ten external datasets, highlighting its importance for general-purpose segmentation.

We further ablated three core architectural components of PAM (Fig. 6e–g, Supplementary Tables S20–S21): (1) Guiding prompt generation: Replacing the Box2Mask-generated prompts with alternatives (Full-box or Adaptive masks) led to only coarse predictions (as shown in Supplementary Fig. S12), whereas Box2Mask provided more structured and accurate guidance, approaching the performance of manual masks. (2) Image–mask encoder design: When the mask encoder was replaced with a shared image encoder, performance dropped substantially, confirming the necessity of decoupling the two encoders to capture complementary information. (3) Multi-scale prompt-guided decoder: Incorporating additional scales progressively improved performance, with the largest gain observed at the second scale (R2). While higher scales (R3, R4) provided marginal further improvements, they also increased computational cost, suggesting potential for future dynamic scale selection strategies.

Overall, these ablation experiments demonstrate that both the dynamic normalization and the architectural innovations—prompt propagation, encoder decoupling, and multi-scale prompt-guided decoding—are crucial to PAM's effectiveness and robustness across diverse medical objects.

Discussion

We developed PAM (Propagating Anything Model) to address the critical need for efficient and accurate volumetric segmentation across diverse 3D medical imaging modalities. PAM offers several key advantages over existing approaches. Using only a single 2D prompt, it achieves high-quality segmentation of any 3D object across a wide range of medical imaging tasks, significantly outperforming existing state-of-the-art methods without being constrained by predefined object categories or specific modalities. Unlike traditional methods, PAM models the continuous flow of information in 3D structures through an innovative propagation framework, capturing inter-slice relationships and going beyond simple object-specific feature learning. This approach enables PAM to demonstrate superior performance and generalization capabilities, making it a versatile tool for various 3D medical segmentation challenges.

The application of successful natural image processing approaches like SAM to medical imaging often struggles with unique challenges posed by 3D medical data, resulting in performance degradation and limited generalization. This gap can be attributed to several factors, primarily the limited availability of annotated data in medical imaging compared to natural image datasets, and the semantic ambiguity in medical objects, particularly in pathological structures. SAM-like methods typically focus on learning to segment specific objects based on prompts. While this approach is effective for 2D natural images due to the abundance of diverse samples, it faces significant challenges in 3D medical imaging. The combination of limited data and high semantic complexity in medical volumes often causes these models to overfit to a small set of object patterns present in the training data, limiting their ability to generalize to the wide variety of structures and anomalies encountered in clinical practice.

In contrast, PAM addresses these challenges by modeling the continuous flow of information across slices, a unique characteristic of 3D medical structures. This approach allows PAM to learn generalizable inter-slice relationships rather than relying on object-specific features. Consequently, PAM avoids the trap of overfitting to limited object patterns and instead captures the underlying structural and semantic continuity across slices. This novel approach enables PAM to better adapt to the diverse and complex nature of 3D medical objects, demonstrating superior generalization to unseen objects and new imaging modalities, even with limited training data.

At the core of PAM is a novel propagation-based segmentation model that integrates CNN-based local feature extraction with Transformer-based attention mechanisms for modeling long-range dependencies. This hybrid architecture enables PAM to efficiently extract precise local features crucial for accurate medical image segmentation while effectively modeling inter-slice relationships and propagating information throughout the volume. As a result, PAM achieves a balance between model capacity and computational efficiency, resulting in faster inference times compared to purely Transformer-based models. This approach contrasts with other methods that either apply 2D segmentation slice-by-slice (e.g., MedSAM) or attempt to model 3D structures directly at the cost of increased computational demands and limited generalization (e.g., SegVol).

Our experimental results demonstrate PAM's superior performance across a wide range of datasets and imaging modalities. By focusing on learning inter-slice relationships, PAM excels at capturing and propagating intricate structural information, enabling it to effectively segment irregular objects. This capability is particularly valuable for complex anatomical

structures and pathological features that deviate from typical shapes. The model's effectiveness in handling such challenging cases underscores its potential to significantly improve medical image analysis in various clinical applications.

Despite these promising results, achieving general volumetric segmentation in medical imaging remains challenging. The diversity of objects and modalities in medical imaging presents a significant hurdle, with variability in object shapes, sizes, and contrasts across different imaging techniques. Additionally, the limited availability of large-scale, annotated 3D medical imaging datasets poses challenges for training and evaluation. These factors highlight the importance of developing robust and adaptable models like PAM that can generalize across diverse medical imaging scenarios.

To address these ongoing challenges and further improve PAM's capabilities, we propose several directions for future work. First, evaluating PAM's impact on downstream clinical tasks by integrating it into clinical workflows and assessing its practical implications for patient diagnosis and prognosis will be crucial. Exploring additional interactive input methods, such as point-based or line-based prompts, could enhance flexibility and user experience. Investigating mechanisms to capture and utilize global contextual information more effectively may improve the segmentation of large or discontinuous structures. Furthermore, exploring the integration of diverse imaging modalities and complementary data types could enhance segmentation accuracy and robustness. Another promising avenue is to develop a dynamic multi-scale prompt-guided decoder, which adaptively selects the appropriate scales of guidance based on the target object's characteristics or the user's precision–efficiency requirements. Such a mechanism could improve segmentation performance while reducing unnecessary computational overhead. Finally, developing strategies for rapid adaptation to new imaging modalities or specific object types with minimal additional training data will be essential for PAM's widespread adoption in varied clinical settings.

In conclusion, PAM represents a significant advancement in 3D medical image segmentation, demonstrating exceptional segmentation capabilities and strong generalization abilities across 44 datasets and multiple medical imaging modalities. By focusing on modeling the continuous flow of information within 3D medical structures, PAM not only achieves superior performance, particularly for complex and irregular objects, but also offers higher inference and interaction efficiencies. These advantages position PAM as a versatile and efficient solution for volumetric segmentation across diverse imaging modalities. Moreover, PAM's success in leveraging the continuous flow of information as a unique learning target opens up new avenues for future research in 3D medical image segmentation. We encourage future research to explore similar innovative learning objectives that capitalize on the inherent characteristics of medical imaging data. Such approaches could include modeling temporal dynamics in 4D imaging, exploring inter-modality information flow, or investigating hierarchical spatial relationships within complex anatomical structures. By identifying and leveraging these unique aspects of medical imaging data, future models may achieve greater performance and generalization capabilities, further advancing the field of medical image segmentation.

Methods

Data acquisition

To enable the training and evaluation of a general-purpose 3D medical image segmentation model, we curated a large-scale benchmark comprising 44 publicly available datasets across four imaging modalities: CT, MRI, PET-CT, and SRX (Fig. 1a and Supplementary Table S1). These datasets, widely adopted in prior segmentation studies, provide high-quality expert annotations for a diverse range of anatomical structures and pathological targets.

To systematically characterize the dataset, we assessed five key attributes for each dataset: total number of 3D scans, voxel count, size anisotropy, spacing anisotropy, and diversity of object types (Supplementary Table S4; Fig. 3b). Following nnUNet[1], we define size anisotropy as the ratio between the smallest and largest physical dimensions of the scans, and spacing anisotropy as the ratio between the smallest and largest voxel spacing values. In accordance with the MedSAM[18] protocol, the 44 datasets were partitioned into 34 internal datasets (D01–D34) for training and validation, and 10 external datasets (D35–D44) for independent testing (Supplementary Tables S2–S3).

Overall, the full dataset covers 168 object categories and contains 1,645,871 annotated 3D instances, spanning a wide spectrum of organs, lesions, and tissues (Fig. 3 and Supplementary Tables S2–S3). This diversity provides a robust foundation for training and evaluating segmentation models under varying anatomical, structural, and imaging conditions, and is critical for ensuring the scalability and generalizability of PAM.

Network architecture

As illustrated in Fig. 2, the proposed propagation-based model for segmenting any 3D medical objects (PAM) is designed to take user-provided 2D prompts together with the corresponding 3D medical image, and to generate full volumetric segmentation of the target object. Specifically, a user first loads the volumetric image and identifies the object of interest in one selected view (e.g., coronal, sagittal, or axial plane). The user then provides a 2D prompt on a guiding slice, as shown in Fig. 2a. If the prompt is a 2D bounding box, PAM invokes the prompt conversion module (called Box2Mask) to transform the region of interest within the box into a binary mask. If the prompt is a freehand contour sketch, PAM applies morphological operations to fill the enclosed region and generate the corresponding mask. Once a guiding mask is obtained, the information propagation module (PropMask) models both the structural and textural continuity between the guiding slice and its adjacent slices, and uses the guiding prompt to infer the segmentation on neighboring slices (Fig. 2b). This process is referred to as propagation in PAM. The newly predicted slices are then treated as updated guiding slices, which iteratively serve as the basis for further propagation. This iterative process continues until the boundary of the 3D image is reached or no additional valid segmentations can be produced, thereby yielding a complete volumetric segmentation of the target object (Fig. 2c). The remainder of this section details the workflow of PAM, including the model inputs, the Box2Mask module, the PropMask module, and the inference process.

Model inputs. The PAM model takes two categories of inputs: a 3D medical image to be segmented and a user-provided prompt that specifies the region of interest. The volumetric image can originate from any imaging modality, including but not limited to CT, MRI, PET-CT, or SRX. The user prompt is defined on a single slice within one of the three standard anatomical views (coronal, sagittal, or axial) and provides 2D guidance regarding the object of interest. PAM supports two styles of 2D prompts: PAM-2DBox (bounding boxes) and PAM-2DMask (freehand contour sketches). The bounding-box style requires minimal user interaction and is suitable for rapid annotation, whereas the contour-based style offers more precise guidance and typically results in higher-quality segmentations.

Two-dimensional (2D) Bounding Box Prompt (PAM-2DBox). In this setting, the user specifies a bounding box enclosing the object of interest on a single slice of the 3D medical image. This type of prompt is widely adopted in general-purpose segmentation models such as MedSAM and SegVol. Unlike these models, which typically require multi-view inputs—i.e., bounding boxes provided on at least two orthogonal views (e.g., axial and sagittal) to construct a 3D bounding box (Supplementary Fig. S2b)—PAM only requires a single-view 2D bounding box (Supplementary Fig. S2a). Since bounding boxes do not directly encode pixel-level spatial information, PAM utilizes the Box2Mask module to convert the input bounding box into a binary mask, thereby aligning it with the internal mask-based processing pipeline.

Two-dimensional (2D) Contour Prompt (PAM-2DMask). In this setting, the user delineates the contour of the object of interest on a single slice. This prompt inherently provides a binary mask and therefore does not require further conversion. It is particularly valuable for objects with irregular

shapes (e.g., stomach or intestines) or with ambiguous boundaries (e.g., tumors or lesion regions).

Box2Mask module. The Box2Mask module converts the coarse PAM-2DBox prompt into a binary 2D mask, enabling consistent downstream processing. Given a bounding box $\mathbf{T}$ on a guiding slice $\mathbf{X}_i$, the module predicts the foreground region via a U-Net-based convolutional network $\mathcal{M}_{\mathrm{U-Net}}$:

$$\hat{\mathbf{P}} = \mathrm{Box2Mask}\left(\mathbf{X}_i, \mathbf{T}\right).$$

The input region of interest is first cropped, normalized, and augmented before being passed through $\mathcal{M}_{\mathrm{U-Net}}$, which outputs a set of multi-resolution probability maps $\left\{\mathbf{P}_1, \cdots, \mathbf{P}_S\right\}$, where $\mathbf{P}_s \in [0,1]^{H_s \times W_s}$. To facilitate learning across resolutions, deep supervision is applied at each scale using a soft Dice loss:

$$\mathcal{L}_{\mathrm{RoI,\,s}} = 1.0 - \frac{2 \times \sum_{i=1}^{W_s} \sum_{j=1}^{H_s} \mathrm{P}_{sij} \mathrm{Y}_{sij}}{\sum_{i=1}^{W_s} \sum_{j=1}^{H_s} \left(\mathrm{P}_{sij}^2 + \mathrm{Y}_{sij}^2\right)},$$

where Y_{sij} is the downsampled ground truth. The overall training loss is averaged over all S scales:

$$\mathcal{L}_{\mathrm{RoI}} = \frac{1}{S} \sum_{s=1}^{S} \mathcal{L}_{\mathrm{RoI,\,s}}.$$

During inference, the probability maps are upsampled to the original resolution ($H \times W$) via bilinear interpolation $\mathcal{R}(\cdot)$ and aggregated by soft voting:

$$\hat{\mathbf{P}} = \mathrm{argmax}\left(\frac{1}{S} \sum_{s=1}^{S} \mathcal{R}(\mathbf{P}_s, H, W)\right) \in \{0,1\}^{H \times W}.$$

The final output $\hat{\mathbf{P}}$ serves as the 2D mask corresponding to the user-provided bounding box prompt.

PropMask module. The PropMask module reconstructs the volumetric segmentation by propagating user-provided (or generated by the Box2-Mask) 2D guidance across adjacent slices, leveraging spatial continuity in the 3D volume. Given a 2D binary mask $\hat{\mathbf{P}}$ on the i-th guiding slice $\mathbf{X}_{\mathrm{gd},i} \in \{0,1,\ldots,255\}^{H \times W}$, and its corresponding mask $\mathbf{M}_{\mathrm{gd},i} \in \{0,1\}^{H \times W}$, the model predicts the mask for each adjacent slice $\mathbf{X}_{\mathrm{adj},j}$ within a K-slice neighborhood ($j \in \{i-K,\ldots,i+K\} \setminus \{i\}$). A single propagation step is defined as:

$$\hat{\mathbf{M}}_{\mathrm{adj},j} = \mathrm{PropMask}\left(\mathbf{X}_{\mathrm{gd},i}, \mathbf{M}_{\mathrm{gd},i}, \mathbf{X}_{\mathrm{adj},j}\right).$$

PropMask integrates both local textures and long-range structural cues using a hybrid architecture combining convolutional encoders and cross-attention. Specifically, CNN encoders extract multi-scale features from the guiding slice, adjacent slice, and guiding mask:

$\mathbf{I}_{\mathrm{gd},si}$, $\mathbf{I}_{\mathrm{adj},sj}$, and $\mathbf{F}_{\mathrm{gd},si}$, respectively, where $s \in \{1,\ldots,S\}$ denotes the scale. At each scale, features are flattened via $\mathcal{F}(\cdot)$ and passed through a cross-attention mechanism:

$$\mathbf{I}'_{\mathrm{adj},sj} = \mathrm{CrossAttention}(\mathbf{Q},\mathbf{K},\mathbf{V})$$

$$= \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{C_s}}\right)\mathbf{V}$$

$$= \mathrm{softmax}\left(\frac{\alpha\left(\mathcal{F}\left(\mathbf{I}_{\mathrm{adj},sj}\right)\right)\beta\left(\mathcal{F}\left(\mathbf{I}_{\mathrm{gd},si}\right)\right)^{\mathrm{T}}}{\sqrt{C_s}}\right)\gamma\left(\mathcal{F}\left(\mathbf{F}_{\mathrm{gd},si}\right)\right),$$

where α, β, and γ are learnable linear projections producing query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) embeddings, and C_s is the feature dimension at scale s. Finally, the attention-enhanced features $\mathbf{I}'_{\mathrm{adj},sj}$ are reshaped and decoded using a U-Net-style decoder with skip connections to generate the predicted mask $\hat{\mathbf{M}}_{\mathrm{adj},j}$. Similar to Box2Mask, deep supervision is applied across scales, and training is optimized using the soft Dice loss.

Propagation-based inference process of PAM. During inference, PAM first converts the user-provided prompt into a 2D mask, identifying the guiding slice and its associated prompt. The PropMask module is then iteratively applied to propagate this guidance to adjacent slices within the same anatomical view. At each step, multi-resolution predictions are aggregated via soft voting to produce the final mask for each target slice. Bidirectional propagation continues from the initial guiding slice until one of two termination conditions is met: (1) the boundary of the 3D image is reached, or (2) no valid predictions (i.e., area below threshold) are generated in a given direction. As illustrated in Fig. 2c, propagation in opposing directions can be performed in parallel and the full procedure is summarized in Algorithm 1.

Algorithm 1. Propagation-based inference process of PAM.

Input: 3D medical image $\mathbf{X} \in \mathbb{R}^{D \times H \times W}$, user-provided prompt $\mathbf{T}$ at slice i (e.g., axial view)
Output: Final prediction 3D mask $\mathbf{M} \in \{0,1\}^{D \times H \times W}$
Initialization: Set neighborhood size K and minimum propagation area threshold τ
Step 1: Convert user prompt into a 2D mask: $\mathbf{M}_{\mathrm{gd},i} \longleftarrow$ Box 2Mask($\mathbf{X}_{\mathrm{gd},i}$, $\mathbf{T}$)
Step 2: Predict segmentation for the K adjacent slices:
for each slice $j \in \{i-K,\ldots,i+K\} \setminus \{i\}$ **do**
  $\hat{\mathbf{M}}_{\mathrm{adj},j} \longleftarrow$ PropMask($\mathbf{X}_{\mathrm{gd},i}, \mathbf{M}_{\mathrm{gd},i}, \mathbf{X}_{\mathrm{adj},j}$)
end for
Step 3: Iteratively update guiding slices:
while True **do**
  if Area($\hat{\mathbf{M}}_{\mathrm{adj},i-K}$) $> \tau$ and $i-K$ is within the image boundary **then**
    Update guiding slice to $i-K$: $\mathbf{X}_{\mathrm{gd},i} \longleftarrow \mathbf{X}_{\mathrm{adj},i-K}$; $\mathbf{M}_{\mathrm{gd},i} \longleftarrow \hat{\mathbf{M}}_{\mathrm{adj},i-K}$; $i \longleftarrow i-K$
    for each slice $j \in \{i-2K,\ldots,i-K+1\}$ **do**
      $\hat{\mathbf{M}}_{\mathrm{adj},j} \longleftarrow$ PropMask($\mathbf{X}_{\mathrm{gd},i}, \mathbf{M}_{\mathrm{gd},i}, \mathbf{X}_{\mathrm{adj},j}$)
    end for
  else
    Stop downward propagation
  end if
  if Area($\hat{\mathbf{M}}_{\mathrm{adj},i+K}$) $> \tau$ and $i+K$ is within the image boundary **then**
    Update guiding slice to $i+K$: $\mathbf{X}_{\mathrm{gd},i} \longleftarrow \mathbf{X}_{\mathrm{adj},i+K}$; $\mathbf{M}_{\mathrm{gd},i} \longleftarrow \hat{\mathbf{M}}_{\mathrm{adj},i+K}$; $i \longleftarrow i+K$
    for each slice $j \in \{i+K+1,\ldots,i+2K\}$ **do**
      $\hat{\mathbf{M}}_{\mathrm{adj},j} \longleftarrow$ PropMask($\mathbf{X}_{\mathrm{gd},i}, \mathbf{M}_{\mathrm{gd},i}, \mathbf{X}_{\mathrm{adj},j}$)
    end for
  else
    Stop downward propagation
  end if
end while
Step 4: Aggregate all propagated predictions to generate the final 3D segmentation $\mathbf{M}$
return M

Implementation details

Data pre-processing. Our data pre-processing pipeline consists of several steps to prepare the data for the Box2Mask module and the PropMask module.

Bounding box generation for Box2Mask. To obtain bounding boxes for the Box2Mask module, we generated the tightest bounding box on the slice where the corresponding foreground mask annotation contains over 100 pixels (Supplementary Figs. S3–S4). We then randomly adjusted the width and height of the bounding box with a scaling ratio between 1.0 and 1.25 to account for potential deviation in actual usage. These processed bounding boxes were used as training data for the Box2Mask module.

RoI task construction for PropMask. For the PropMask module, we constructed RoI tasks (Supplementary Figs. S6–S7). We generated the tightest bounding box around the mask of the guiding slice and then randomly adjusted its width and height with a scaling ratio between 1.0 and 2.0 to capture the context around the target object. This adjusted bounding box was then used to crop both the guiding slice and the adjacent slices sampled within the propagation thickness, forming the cropped RoI tasks as training data for the PropMask module.

Dynamic image normalization. For Box2Mask, since the foreground region of the RoI image is unknown initially, we first normalize the image by applying a series of candidate min-max values. These candidate min values are based on the 5th to 20th percentiles, and the candidate max values are derived from the 90th to 95th percentiles of the entire RoI image. After performing Box2Mask inference with these candidate settings and merging the resulting predictions, we can obtain an initial estimate of the foreground region. Subsequently, we apply min-max normalization to the RoI image again, using the 5th and 95th percentiles of the pixel distribution within the estimated foreground region. Finally, the Box2Mask module is applied once more to obtain the final prediction.

For PropMask, since the guiding prompt already defines the foreground region, we directly compute the 5th and 95th percentiles of the foreground pixels and apply the min-max normalization accordingly.

Data augmentation. To optimize training efficiency, we applied offline data augmentation for both the Box2Mask and PropMask modules. For the Box2Mask module, we augmented each sample five times. Each image had a 50% chance of being flipped horizontally and vertically. Additionally, we randomly adjusted the image's brightness and contrast, also with a 50% probability, setting the adjustment ranges to [-0.2, 0.2]. The images were also rotated randomly up to 45 degrees with a 50% probability, filling any areas outside the original boundaries with a constant value (typically black).

For the PropMask module, since the fundamental training unit is a task containing several images (typically 20 adjacent images and one guiding image), we applied augmentation to each image within a task. Specifically, each image in a task had a 50% chance of being flipped (either horizontally or vertically) and an independent 50% chance of being rotated up to 45 degrees.

After augmentation, all samples were uniformly resized to a resolution of 224×224 for input into both the Box2Mask and PropMask modules.

Comparison methods and dataset partitioning. In this study, we evaluate our proposed model against three baseline models: MedSAM, SegVol, and nnUNet. These baseline models were selected based on their established performance and general applicability in medical image segmentation tasks. The first two models, MedSAM and SegVol, are general-purpose models that, after being trained on large-scale datasets, can be applied to any object, allowing for a direct comparison without the need for retraining.

MedSAM. For MedSAM, we strictly followed the setup and inference instructions provided in the official repository, available at MedSAM GitHub (https://github.com/bowang-lab/MedSAM.git). We loaded the pretrained weights from the official Google Drive link (https://drive.google.com/drive/folders/1ETWmi4AiniJeWOt6HAsYgTjYv_fkgzoN) as specified by the authors. To ensure a fair and consistent comparison, we adhered to the dataset partitioning provided in Supplementary Tables 1–4 of MedSAM's Supplementary Information (https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-024-44824-z/MediaObjects/41467_2024_44824_MOESM1_ESM.pdf), ensuring that the same internal and external datasets were used for both training and evaluation.

SegVol. For SegVol, we utilized the official open-source code available at SegVol GitHub (https://github.com/BAAI-DCAI/SegVol.git) and loaded the pretrained weights from Hugging Face (https://huggingface.co/BAAI/SegVol/tree/main) for inference testing. Similar to MedSAM, we strictly followed the official configuration settings provided in the repository to maintain consistency with the original setup. This ensured that our use of SegVol was in accordance with the guidelines set forth by its developers, allowing for a fair comparison between models.

nnUNet. We also included nnUNet as a baseline model. nnUNet is a supervised learning model that requires specific training for each dataset, unlike the general-purpose models mentioned earlier. For this study, we evaluated nnUNet's performance on a representative subset comprising ten datasets to provide a comprehensive benchmark. We performed a random split of each dataset into training and validation sets and trained nnUNet on the training sets and performed inference on the validation sets. The training was conducted using the latest version (v2) of nnUNet, which was obtained from the official nnUNet GitHub repository (https://github.com/MIC-DKFZ/nnUNet).

Data partitioning. Following the pre-processing steps described in Section 5.3.1, we obtained a total of 19,344,368 samples for the Box2Mask module and 1,345,871 tasks across 44 datasets. In accordance with MedSAM's data partitioning protocol, we divided these data into internal and external validation datasets. Furthermore, the internal validation dataset was further split into training and validation sets at an 80:20 ratio. The final distribution of samples was as follows: for the Box2Mask module, 14,974,620 samples were used for training, 3,782,206 for internal validation, and 587,542 for external validation (Supplementary Table S2); for PropMask module, 1,020,576 tasks were used for training, 258,889 for internal validation, and 66,406 for external validation (Supplementary Table S3).

We trained the overall PAM on the training set, using the internal validation set to evaluate model performance. The external validation dataset served to demonstrate the robustness of PAM and its zero-shot capability with unseen objects and datasets.

Training configuration. We implemented our models using PyTorch[50] (version 2.0.0) and executed them on a server equipped with the CUDA platform (version 11.8). Both the Box2Mask module and the PropMask module were trained using four NVIDIA A800-SXM4-80GB GPUs and 64 Intel(R) Xeon(R) Platinum 8358 P CPUs (2.60 GHz). We utilized the AdamW optimizer with an initial learning rate of 1e-3 for the Box2Mask module and 5e-4 for the PropMask module, as well as a weight decay of 1e-4. The learning rate was adjusted according to the Cosine Annealing LR schedule with a maximum period of 100 epochs and a minimum eta of 1e-5.

For the Box2Mask module, during each epoch, we randomly selected 10,000 samples for training and conducted evaluations every 20 epochs using a set of 5000 randomly sampled validation samples. The training lasted for 4100 epochs, with a batch size of 1024, over a span of about six days. Supplementary Fig. S5 illustrates the training and validation curves.

We selected the latest checkpoint as the final weight configuration for our Box2Mask module.

For the PropMask module, throughout the training process, we randomly selected 10,000 tasks per epoch. Each task consists of the guiding slice and four randomly sampled adjacent slices. Evaluations were conducted every 20 epochs using a set of 5000 randomly selected validation tasks. The training extended over 4500 epochs, with a batch size of 160, lasting ~7 days. Supplementary Fig. S8 displays the training and validation loss curves. We chose the most recent checkpoint as the final weight configuration for our PropMask module.

Evaluation metrics. For segmentation performance evaluation, six commonly used metrics were employed:

Dice Similarity Coefficient (DSC):. DSC is a set similarity metric that measures the overlap between the predicted segmentation and the ground truth, with values ranging from 0.0 to 1.0. A higher DSC indicates better segmentation quality.

$$\mathrm{DSC} = \frac{2|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A}| + |\mathbf{B}|}$$

where $\mathbf{A}$ represents the ground truth mask, $\mathbf{A}$ the predicted mask, and $|\cdot|$ the cardinality of a set.

Hausdorff Distance at the 95th percentile (HD95):. HD95 measures the boundary discrepancy between the predicted and ground truth masks by computing the 95th percentile of the bidirectional Hausdorff distance. A lower HD95 indicates closer boundary alignment and is less sensitive to outliers than the maximum Hausdorff distance.

Average Surface Distance (ASD):. ASD quantifies the mean distance between the surfaces of the predicted and ground truth segmentations in both directions. A lower ASD indicates a more accurate boundary approximation.

Normalized Surface Dice (NSD):. NSD evaluates the proportion of surface points on the predicted mask that lie within a tolerance distance δ from the ground truth surface. This metric provides a clinically meaningful measure of segmentation quality, as it accounts for the acceptable boundary deviation in practical applications.

Sensitivity:. Sensitivity measures the proportion of ground truth positives (foreground voxels) that are correctly identified by the predicted segmentation. A higher sensitivity indicates fewer false negatives and better detection of the target region.

Specificity:. Specificity measures the proportion of ground truth negatives (background voxels) that are correctly identified by the predicted segmentation. A higher specificity indicates fewer false positives and better discrimination between foreground and background.

For efficiency evaluation, we measured inference time, defined as the elapsed time from reading input samples to writing final segmentation results. Lower inference time indicates higher computational efficiency.

Additionally, to quantify the complexity and regularity of segmentation targets, we employed three geometric metrics:

Box Ratio:. This metric evaluates the similarity between the target mask and its tight bounding box, defined as:

$$\mathrm{BoxRatio}(\mathrm{M}) = \frac{|\mathrm{M}|}{|\mathrm{Box}(\mathrm{M})|},$$

where $\mathrm{Box}(\mathbf{M})$ is the tightest bounding box around mask $\mathbf{M}$, and $|\cdot|$ represents the area measured in pixels.

Convex Ratio:. This measure quantifies how convex the target mask is, expressed as:

$$\mathrm{ConvexRatio}(\mathrm{M}) = \frac{|\mathrm{M}|}{|\mathrm{ConvexHull}(\mathrm{M})|},$$

where $\mathrm{ConvexHull}(\mathbf{M})$ is the convex hull of mask $\mathbf{M}$.

Inverse Rotational Inertia (IRI):. This metric assesses how spread out the area of the target mask is, calculated as:

$$\mathrm{IRI}(\mathrm{M}) = \frac{0.75|\mathrm{M}|}{\left(0.8\pi \mathrm{RI}(\mathrm{M})^{\frac{5}{3}}\right)},$$

where the rotational inertia of $\mathbf{M}$ relative to its centroid $c_{\mathbf{M}}$ is $\mathrm{RI}(\mathrm{M}) = \sum_{x\in\mathrm{M}} ||x - c_{\mathrm{M}}||_2$, with x being the coordinate of each pixel in the mask, and $c_{\mathbf{M}}$ being the centroid coordinate.

Statistics and reproducibility

Sample sizes and the number of datasets were determined based on the availability of all publicly accessible volumetric segmentation datasets that we could download and process. No statistical method was used to predetermine the sample sizes or the number of datasets.

For performance evaluation and statistical analysis:

Performance metrics, including DSC, boundary-based measures, sensitivity, and specificity, were calculated for every object type in each dataset, with the dataset-level performance obtained by averaging over its constituent objects.

For paired comparisons of different methods across multiple datasets, we applied a paired *t*-test when the distributional assumptions were satisfied; otherwise, the Wilcoxon signed-rank test was used. A *P*-value less than 0.05 indicated significant differences.

Performance comparisons across multiple experimental groups in ablation studies were conducted using a one-way ANOVA test. A *P*-value greater than 0.05 indicated no significant difference in performance across the groups, suggesting stability in the experimental results.

The relationship between model performance improvements and object regularity was assessed using the Pearson correlation coefficient. The correlation coefficient (r) quantifies the strength and direction of the linear association between the two variables, where a negative r indicates that higher irregularity is associated with greater model improvements. The corresponding *P*-value (*P*) measures the statistical significance of the observed correlation, with smaller values indicating stronger evidence against the null hypothesis of no correlation.

We used R (version 4.1.3) for results analysis and statistical analyses, and Python (version 3.7.10) for model construction, training, and inference. To ensure reproducibility, we provide detailed configurations of our methodology, visual demonstrations, and experimental results in the supplementary materials. These include Supplementary Texts, Supplementary Figs. S1–S12, and Supplementary Tables S1–S22. All procedures were conducted in accordance with good clinical practice and data privacy regulations.

Data availability

All datasets referenced in this study are publicly available. Supplementary Table S1 provides the download links. The source code and supporting materials are available in the Supplementary Materials. All R packages employed in this study can be found on CRAN(https://cran.r-project.org/web/packages/available_packages_by_name.html) or Bioconductor(https://www.bioconductor.org/).

Code availability

The PyTorch implementation of PAM and pretrained checkpoints for Box2Mask and PropMask is released on GitHub (https://github.com/czifan/PAM).

Received: 19 June 2025; Accepted: 11 October 2025;
Published online: 02 December 2025

References

1. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
2. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
3. Cao, K. et al. Large-scale pancreatic cancer detection via non-contrast CT and deep learning. *Nat. Med.* **29**, 3033–3043 (2023).
4. Yuan, M. et al. Devil is in the queries: advancing mask transformers for real-world medical image segmentation and out-of-distribution localization. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23879−23889 (IEEE, 2023).
5. Hu, Y. et al. AI-based diagnosis of acute aortic syndrome from noncontrast CT. *Nat. Med.* https://doi.org/10.1038/s41591-025-03916-z (2025).
6. Zhou, W. et al. Multimodal model for the diagnosis of biliary atresia based on sonographic images and clinical parameters. *NPJ Digit Med.* **8**, 371 (2025).
7. Ferrari, V. et al. Value of multidetector computed tomography image segmentation for preoperative planning in general surgery. *Surg. Endosc.* **26**, 616–626 (2012).
8. He, M. et al. Associations of subcutaneous fat area and Systemic Immune-inflammation Index with survival in patients with advanced gastric cancer receiving dual PD-1 and HER2 blockade. *J. Immunother. Cancer* **11**, e007054 (2023).
9. He, M. et al. Deep learning model based on multi-lesion and time series CT images for predicting the benefits from anti-HER2 targeted therapy in stage IV gastric cancer. *Insights Imag.* **15**, 59 (2024).
10. Li, J. et al. CT-based delta radiomics in predicting the prognosis of stage IV gastric cancer to immune checkpoint inhibitors. *Front Oncol.* **12**, 1059874 (2022).
11. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
12. Scott, A. J. et al. Rewiring of cortical glucose metabolism fuels human brain cancer growth. *medRxiv* **646**, 413−422 (2023).
13. Bao, P., Wang, G., Yang, R. & Dong, B. Deep reinforcement learning for beam angle optimization of intensity-modulated radiation therapy. *arXiv* https://doi.org/10.48550/arXiv.2303.03812 (2023).
14. Chen, Z. et al. Predicting gastric cancer response to anti-HER2 therapy or anti-HER2 combined immunotherapy based on multi-modal data. *Signal Transduct. Target Ther.* **9**, 222 (2024).
15. Zaidi, H. & El Naqa, I. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. *Eur. J. Nucl. Med. Mol. Imaging* **37**, 2165–2187 (2010).
16. Chen, T. et al. Whole slide image based deep learning refines prognosis and therapeutic response evaluation in lung adenocarcinoma. *NPJ Digit Med.* **8**, 69 (2025).
17. Lu, L., Dercle, L., Zhao, B. & Schwartz, L. H. Deep learning for the prediction of early on-treatment response in metastatic colorectal cancer from serial medical imaging. *Nat. Commun.* **12**, 6654 (2021).
18. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
19. Wang, G. et al. DeepIGeoS: A deep interactive geodesic framework for medical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 1559–1572 (2019).
20. Wang, S. et al. Annotation-efficient deep learning for automatic medical image segmentation. *Nat. Commun.* **12**, 5915 (2021).
21. Dorent, R. et al. CrossMoDA 2021 challenge: benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. *Med. Image Anal.* **83**, 102628 (2023).
22. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference On Medical Image Computing And Computer-Assisted Intervention* 234−241 (Springer, 2015).
23. Zhang, J. et al. Rapid vessel segmentation and reconstruction of head and neck angiograms from MR vessel wall images. *NPJ Digit Med.* **8**, 483 (2025).
24. Wang, S. et al. GH-UNet: group-wise hybrid convolution-VIT for robust medical image segmentation. *NPJ Digit Med.* **8**, 426 (2025).
25. Soomro, T. A. et al. Image segmentation for MR brain tumor detection using machine learning: a review. *IEEE Rev. Biomed. Eng.* **16**, 70–90 (2023).
26. Xie, W., Jacobs, C., Charbonnier, J. P. & van Ginneken, B. Relational modeling for robust and efficient pulmonary lobe segmentation in CT Scans. *IEEE Trans. Med. Imaging* **39**, 2664–2675 (2020).
27. Primakov, S. P. et al. Automated detection and segmentation of non-small cell lung cancer computed tomography images. *Nat. Commun.* **13**, 3423 (2022).
28. Tang, X. et al. Whole liver segmentation based on deep learning and manual adjustment for clinical use in SIRT. *Eur. J. Nucl. Med. Mol. Imaging* **47**, 2742–2752 (2020).
29. Antonelli, M. et al. The medical segmentation decathlon. *Nat. Commun.* **13**, 4128 (2022).
30. Ravi, N. et al. SAM 2: Segment anything in images and videos. In *Proc. International Conference on Representation Learning,* 28085–28128 (2025).
31. Kirillov, A. et al. Segment anything. In *Proc. IEEE/CVF International Conference on Computer Vision,* 4015-4026 (IEEE, 2023).
32. Zhu, J., Hamdi, A., Qi, Y., Jin, Y. & Wu, J. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv* https://doi.org/10.48550/arXiv.2408.00874 (2024).
33. Wu, J. et al. Medical SAM adapter: adapting segment anything model for medical image segmentation. *Med Image Anal.* **102**, 103547 (2025).
34. Zhou, T., Zhang, Y., Zhou, Y., Wu, Y. & Gong, C. Can Sam segment polyps? *arXiv* https://doi.org/10.48550/arXiv.2304.07583 (2023).
35. Wald, T. et al. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv* https://doi.org/10.48550/arXiv.2304.05396 (2023).
36. He, S., Bao, R., Li, J., Grant, P. E. & Ou, Y. Accuracy of segment-anything model (SAM) in medical image segmentation tasks. *arXiv* https://arxiv.org/pdf/2304.09324 (2023).
37. Hu, C., Xia, T., Ju, S. & Li, X. When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation. *arXiv* https://doi.org/10.48550/arXiv.2304.08506 (2023).
38. Deng, R. et al. Segment anything model (SAM) for digital pathology: assess zero-shot segmentation on whole slide imaging. *IS&T Int. Symp. Electron Imag.* **37**, COIMG-132 (2025).
39. Mazurowski, M. A. et al. Segment anything model for medical image analysis: an experimental study. *Med. Image Anal.* **89**, 102918 (2023).
40. Huang, Y. et al. Segment anything model for medical images? *Med. Image Anal.* **92**, 103061 (2024).
41. Du, Y., Bai, F., Huang, T. & Zhao, B. Segvol: Universal and interactive volumetric medical image segmentation. *Adv. Neural Inf. Process. Syst.* **37**, 110746–110783 (2024).
42. Chao, C. J. et al. Foundation versus domain-specific models for left ventricular segmentation on cardiac ultrasound. *NPJ Digit Med.* **8**, 341 (2025).
43. Zhang, Y. et al. Generalist medical foundation model improves prostate cancer segmentation from multimodal MRI images. *NPJ Digit Med.* **8**, 372 (2025).

44. Zhao, Z. et al. Large-vocabulary segmentation for medical images with text prompts. *NPJ Digit Med.* **8**, 566 (2025).
45. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M. & Zhong, J. Attention is all you need in speech separation. In *ICASSP* 2021–2021 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 21−25 (IEEE, 2021).
46. Zhao, T. et al. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. *arXiv* preprint arXiv:2405.12971 (2024).
47. Eisenhauer, E. A. et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
48. Xie, W., Willems, N., Patil, S., Li, Y. & Kumar, M. Sam fewshot finetuning for anatomical segmentation in medical images. In *Proc. IEEE/CVF Winter Conference On Applications Of Computer Vision* 3253−3261 (IEEE, 2024).
49. Li, Y., Hu, M. & Yang, X. Polyp-sam: Transfer sam for polyp segmentation. In *Medical Imaging 2024: Computer-Aided Diagnosis*, 749−754 (SPIE, 2024).
50. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. In *Proc. 33rd Conference on Neural Information Processing Systems,* 8026−8037 (NeurIPS, 2019).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (72573036 to Li Zhang, 12090022 to Bin Dong, U24A20759 to Lei Tang, 82441018), Clinical Medicine Plus X-Young Scholars Project of Peking University (PKU2023LCXQ041 to Li Zhang), Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0520600). Bin Dong is supported by the New Cornerstone Investigator Program. This work is supported by Biomedical Computing Platform of National Biomedical Imaging Center, Peking University. The authors sincerely thank all the participants involved in this study for their contribution.

Author contributions

Zifan Chen, Lei Tang, Li Zhang, and Bin Dong designed the study. Zifan Chen, Xinyu Nan, Jiazheng Li, Jie Zhao, Haifeng Li, Ziling Lin, Haoshen Li, Heyun Chen, and Yiting Liu collected and processed the data. Zifan Chen, Li Zhang and Bin Dong proposed the model design. Zifan Chen and Xinyu Nan implemented the code, trained the model, and constructed the experimental workflow. Zifan Chen, Xinyu Nan, Jie Zhao, HaiFeng Li, Ziling Lin, Haoshen Li, and Heyun Chen jointly conducted inference of the model across multiple datasets. Jie Zhao, Haifeng Liu, Ziling Lin, Haoshen Li, and Heyun Chen reviewed and validated the code. Jiazheng Li performed the interactive time analysis. Jiazheng Li, Yiting Liu, and Lei Tang evaluated the segmentation results of medical objects from the clinical perspective. Zifan Chen, Xinyu Nan, and Jiazheng Li drafted the manuscript, prepared the figures, and organized the tables. Lei Tang, Li Zhang, and Bin Dong supervised the project and revised the manuscript. Zifan Chen, Xinyu Nan, and Jiazheng Li contributed equally to this work. All co–first authors and corresponding authors directly accessed and verified the underlying data reported in the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-02087-y.

Correspondence and requests for materials should be addressed to Lei Tang, Li Zhang or Bin Dong.