nature methods

Article

https://doi.org/10.1038/s41592-024-02499-w

A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities

Received: 21 May 2024

Accepted: 2 October 2024

Published online: 18 November 2024

Check for updates

Theodore Zhao[1,6], Yu Gu[1,6], Jianwei Yang[1], Naoto Usuyama[1], Ho Hin Lee[1], Sid Kiblawi[1], Tristan Naumann[1], Jianfeng Gao[1], Angela Crabtree[3], Jacob Abel[2], Christine Moung-Wen[2], Brian Piening[2,3], Carlo Bifulco[2,3], Mu Wei[1] ✉, Hoifung Poon[1] ✉ & Sheng Wang[4,5] ✉

Biomedical image analysis is fundamental for biomedical discovery. Holistic image analysis comprises interdependent subtasks such as segmentation, detection and recognition, which are tackled separately by traditional approaches. Here, we propose BiomedParse, a biomedical foundation model that can jointly conduct segmentation, detection and recognition across nine imaging modalities. This joint learning improves the accuracy for individual tasks and enables new applications such as segmenting all relevant objects in an image through a textual description. To train BiomedParse, we created a large dataset comprising over 6 million triples of image, segmentation mask and textual description by leveraging natural language labels or descriptions accompanying existing datasets. We showed that BiomedParse outperformed existing methods on image segmentation across nine imaging modalities, with larger improvement on objects with irregular shapes. We further showed that BiomedParse can simultaneously segment and label all objects in an image. In summary, BiomedParse is an all-in-one tool for biomedical image analysis on all major image modalities, paving the path for efficient and accurate image-based biomedical discovery.

Biomedical image analysis is critical to biomedical discovery because imaging is one of the most important tools for studying physiology, anatomy and function at multiple scales from the organelle level to the organ level[1–4]. Holistic image analysis comprises multiple subtasks, such as segmentation, detection and recognition of biomedical objects. Segmentation aims to divide an image into segments representing different objects, often requiring the aid of a user-provided bounding box for each object of interest[5,6]. Detection aims to identify the location of an object of interest in the image[7], whereas recognition aims to identify all objects within an image[8]. Standard image analysis methods typically approach these tasks separately, using specialized tools for individual tasks[9]. Despite their encouraging performance, such a disjointed approach misses opportunities for joint learning and reasoning across these interdependent tasks.

For example, a lot of previous image analysis works focus on segmentation alone, thus ignoring key semantic information from interdependent tasks, such as metadata and object type names. This results in suboptimal segmentation while imposing substantial burden on users, as many state-of-the-art segmentation tools require users to provide a tight bounding box indicating the location of an object of interest[10,11].

[1]Microsoft Research, Redmond, WA, USA. [2]Providence Genomics, Portland, OR, USA. [3]Earle A. Chiles Research Institute, Providence Cancer Institute, Portland, OR, USA. [4]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. [5]Department of Surgery, University of Washington, Seattle, WA, USA. [6]These authors contributed equally: Theodore Zhao, Yu Gu. ✉e-mail: muhsin.wei@microsoft.com; hoifung@microsoft.com; swang@cs.washington.edu

https://doi.org/10.1038/s41592-024-02499-w

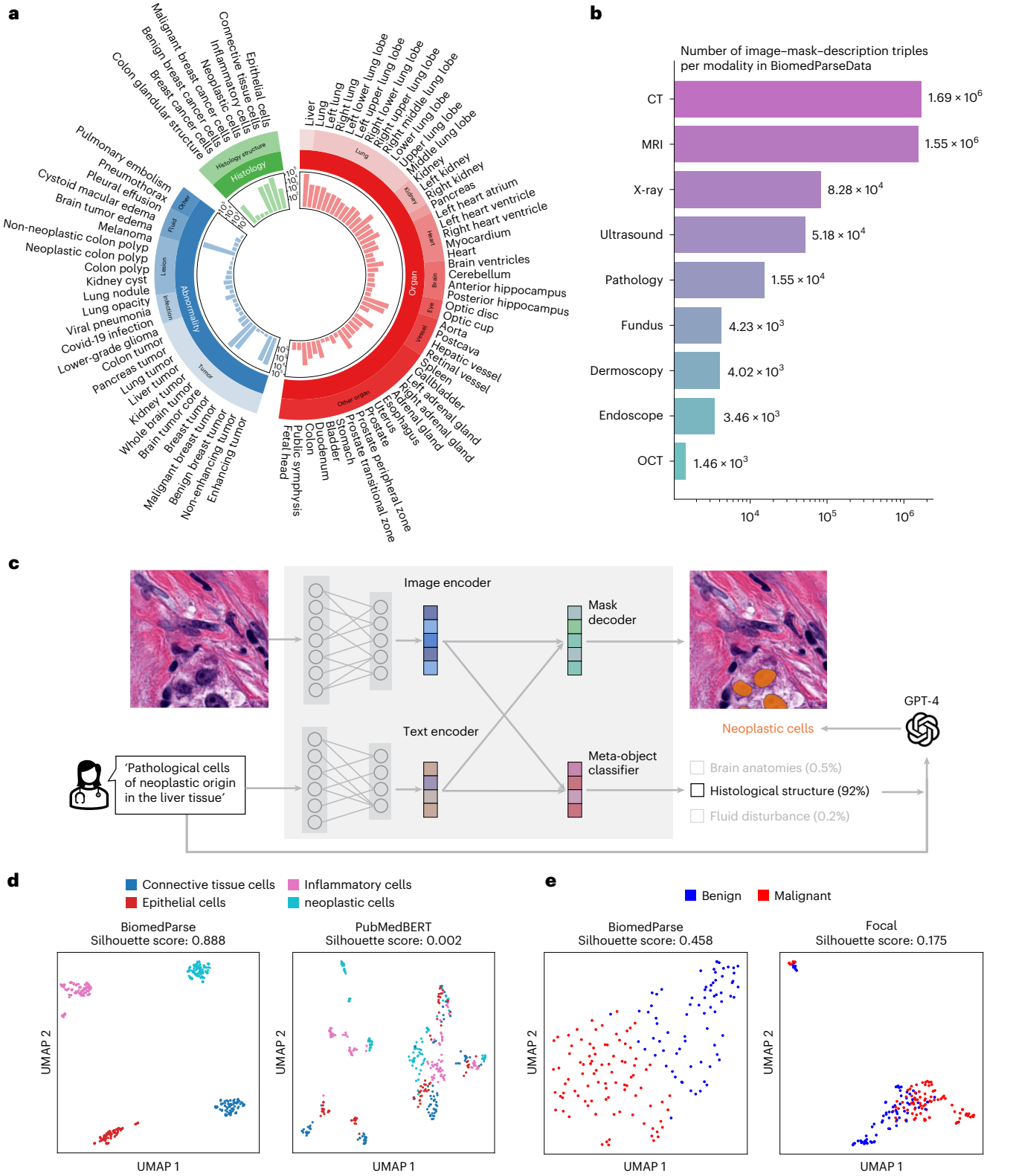


Fig. 1 | Overview of BiomedParse and BiomedParseData. a, The GPT-4 constructed ontology showing a hierarchy of object types that are used to unify semantic concepts across datasets. Bar plots showing the number of images containing that object type. **b**, Bar plot showing the number of image–mask–description triples for each modality in BiomedParseData. CT, computed tomography; MRI, magnetic resonance imaging; OCT, optical coherence tomography. **c**, Flowchart of BiomedParse. BiomedParse takes an image and a text prompt as input and then outputs the segmentation masks for the objects specified in the prompt. Image-specific manual interaction such as bounding box or clicks is not required in our framework. To facilitate semantic learning for the image encoder, BiomedParse also incorporates a learning objective to classify the meta-object type. For online inference, GPT-4 is used to resolve text prompt into object types using the object ontology, which also uses the meta-object type output from BiomedParse to narrow down candidate semantic labels. **d**, Uniform Manifold Approximation and Projection (UMAP) plots contrasting the text embeddings for different cell types derived from BiomedParse text encoder (left) and PubMedBERT (right). **e**, UMAP plots contrasting the image embeddings for different cell types derived from BiomedParse image encoder (left) and Focal (right).

The bounding box requirement leads to three limitations. First, users have to manually draw bounding boxes in the image, which requires domain expertise to identify the locations and shapes of the target objects. Second, bounding boxes, which are often rectangular, fall short of accurately representing objects with irregular or complex shapes. Third, bounding box-based approaches are not scalable for images containing a large number of objects, such as segmenting cells in a whole-slide pathology image, as users need to provide a bounding box for each object.

In this paper, we propose to approach biomedical image analysis as image parsing, a unifying framework for joint learning and reasoning across segmentation, detection and recognition[12–14]. Specifically, we have developed BiomedParse, a biomedical foundation model for image analysis that is capable of carrying out all three tasks by leveraging their interdependencies, thus addressing key limitations in traditional methods. In particular, joint learning of object detection and recognition eliminates the need for user-specified bounding boxes, as segmentation can be carried out using semantic labels from text prompts alone.

The major bottleneck for pretraining BiomedParse is data. While biomedical segmentation datasets abound[15–18], there are relatively few previous works on object detection and recognition in biomedicine, let alone datasets covering all three tasks. To address this problem, we propose a new approach for pretraining BiomedParse using no more than standard segmentation datasets. The key insight is to leverage readily available natural language labels or descriptions accompanying those datasets and use GPT-4 to harmonize these noisy, unstructured texts with established biomedical object ontologies. This enables us to construct BiomedParseData, a biomedical image analysis dataset comprising 3.4 million triples of image, segmentation mask and semantic labels of the biomedical object and 6.8 million image–mask–description triples, from over 1 million images. The semantic labels encompass 82 major biomedical object types across 9 imaging modalities.

Unlike segmentation methods that focus on identifying salient segment boundary within a bounding box, BiomedParse learns to model typical shape of each object class, thus mimicking how humans perceive objects in an image. BiomedParse can segment images using text prompts alone (for example 'inflammatory cells in breast pathology'), without requiring any user-specified localization such as bounding boxes. Consequently, BiomedParse can better recognize and segment objects of irregular and complex shapes, which are very challenging for traditional methods using rectangular bounding boxes. Moreover, BiomedParse can recognize all objects in an image, without requiring any user text prompt.

We conduct a large-scale study to evaluate BiomedParse on 102,855 held-out image–mask-label triples across nine modalities for segmentation, detection and recognition. On segmentation, BiomedParse established new state-of-the-art results, outperforming previous best methods such as MedSAM[11] and SAM[10]. Moreover, using text prompts alone, BiomedParse is much more scalable than these previous methods, which require orders of magnitude more user operations in specifying object-specific bounding boxes to perform competitively. We also demonstrated that BiomedParse can accurately detect invalid text prompts describing nonexistent objects in the image. BiomedParse achieves even larger improvement in image analysis accuracy for irregular-shaped objects, attaining a 0.857 Dice score that is 39.6% higher than the best-competing method. On recognition, we show how BiomedParse can accurately segment and label all objects without any user-specified text prompt. Collectively, we introduce a biomedical foundation model for biomedical image analysis, achieving superior performance on segmentation, detection and recognition, paving the way for large-scale image-based biomedical discovery.

Results

Overview of BiomedParse and BiomedParseData

To develop a model that can jointly conduct segmentation, detection and recognition, we need a supervision dataset that covers all three tasks. To the best of our knowledge, no such datasets exist. To this end, we created the dataset BiomedParseData by combining 45 biomedical image segmentation datasets and using GPT-4 to generate the canonical semantic label for each segmented object.

The key insight is that existing segmentation datasets often contain valuable semantic information about the segmented objects; however, such information typically resides in noisy and inconsistent natural language text descriptions that do not conform to standard biomedical ontologies. To address this challenge, we use GPT-4 to create a unifying biomedical object ontology for image analysis and harmonize natural language descriptions with this ontology (Methods). This ontology encompasses three main categories (histology, organ and abnormality), 15 meta-object types and 82 specific object types (Fig. 1a). The resulting BiomedParseData contains 3.4 million distinct image–mask–label triples, spanning nine imaging modalities and 25 anatomic sites (Fig. 1b and Extended Data Fig. 1), representing a large-scale and diverse dataset for semantic-based biomedical image analysis.

To make BiomedParse better equipped in handling diverse text prompts not covered by the canonical semantic labels, we also use GPT-4 to synthesize synonymous text descriptions for each semantic label and sample from them during training (Methods and Supplementary Figs. 1 and 2). This yielded a total of 6.8 million image–mask–description triples.

While our method does not use bounding boxes, previous state-of-the-art methods such as MedSAM and SAM generally require prespecified bounding boxes. We consider two scenarios to provide the bounding boxes: oracle bounding box (the minimum rectangular bounding box covering a segmented object) and bounding box created by Grounding DINO[19], a state-of-the-art object detection method that can generate bounding boxes from text prompt of an object label. Grounding DINO does not perform segmentation.

BiomedParse adopts a modular design under the SEEM architecture[20], comprising an image encoder (for encoding the input image), a text encoder (for encoding the text prompt), a mask decoder (for outputting the segmentation mask) and a meta-object classifier (for joint training of image encoder with object semantics) (Fig. 1c). The image and text encoders were initialized using state-of-the-art Focal[21] and PubMedBERT[22], respectively.

Before evaluating image analysis results, we first examine the quality of embeddings derived from BiomedParse. Specifically, we compare the text embeddings from BiomedParse to those from PubMedBERT. We found that embeddings from BiomedParse can better distinguish fine-grained cell types, with a Silhouette score of 0.89, which is much higher than using the embeddings from PubMedBERT (Fig. 1d and Extended Data Fig. 2). We also compare the image embeddings from BiomedParse with those from Focal. We observed that embeddings from BiomedParse are more predictive of tumor malignancy on a pathology dataset[23] (Fig. 1e). The superior performance of the text and image embeddings from BiomedParse necessitates the training of BiomedParse using BiomedParseData, raising our confidence that BiomedParse can be an effective approach for biomedical image analysis.

Accurate and scalable segmentation across nine modalities

We first evaluated BiomedParse on biomedical image segmentation using the held-out set comprising 102,855 test instances (image–mask–label triples) across nine imaging modalities (Fig. 2a and Extended Data Figs. 2 and 3). We observed that BiomedParse achieved the best Dice score, even against the best-competing method MedSAM with the oracle bounding box as input (paired t-test P value $<10^{-4}$). In the more realistic setting when MedSAM or SAM is supplied with bounding boxes generated by Grounding DINO, the superiority of BiomedParse is even more prominent in end-to-end biomedical object detection and segmentation, especially in more challenging modalities such as pathology and computed tomography (CT) where irregular-shaped objects abound. By training on domain-specific datasets, both BiomedParse and

MedSAM outperform general-domain methods such as SAM. We further observed that BiomedParse outperformed other text prompt-based approaches (Supplementary Fig. 3 and Extended Data Fig. 4), including SEEM[20], SegVol[24] and SAT[25], task-specific segmentation method CellViT[26] (Supplementary Fig. 4) on cell segmentation, widely used supervised method Swin UNETR[27] (Supplementary Fig. 5), nnU-Net[28] and DeepLabV3+[29] (Supplementary Fig. 6) and universal biomedical segmentation model UniverSeg[30] (Supplementary Fig. 7). BiomedParse also outperformed SAM continually trained on BiomedParseData and MedSAM continually trained on BiomedParseData (Extended Data Fig. 5), even though both SAM and MedSAM utilized oracle bounding boxes for training and inference.

We showed examples comparing BiomedParse segmentation and the ground truth across multiple imaging modalities, demonstrating the generalizability of BiomedParse (Fig. 2b). We further compared BiomedParse on a benchmark created by MedSAM[11] encompassing 50 tasks and again observed the best performance by BiomedParse, even against MedSAM with oracle bounding box (paired t-test P value $<10^{-2}$), further demonstrating the superiority of BiomedParse (Extended Data Fig. 6). In addition to being more accurate, BiomedParse is more scalable compared to bounding box-based approaches, which stems from the generalizability of text prompts across images of the same modality or anatomical site, thus eliminating the need for laborious user operations in providing a tight bounding box for each object. To demonstrate this, we compared BiomedParse and previous state-of-the-art methods MedSAM and SAM on a cell segmentation dataset with 42 colon pathology images (Fig. 2c). Using a single text prompt 'glandular structure in colon pathology image', BiomedParse achieves a 0.942 median Dice score, whereas neither SAM nor MedSAM achieves a median Dice score higher than 0.75 without tight bounding boxes as input. In fact, to achieve competitive results comparable to BiomedParse with a single text prompt, MedSAM requires the users to supply a tight bounding box for each of the 430 cells in these images (Fig. 2c). In general, our results reveal that the bounding box-based approach is much less accurate on irregular-shaped objects, such as tumors and abnormal cells (Fig. 2d,e). In contrast, BiomedParse still attained highly accurate segmentation for such objects. The scalability and accuracy of BiomedParse bode well for its utility in real-world applications.

BiomedParse can also detect invalid text prompts (for example, the request to identify a brain tissue in a chest X-ray image), by calculating a P value using Kolmogorov–Smirnov (K–S) test (Methods). From preliminary experiments, we found that invalid text prompts have an average K–S test P value smaller than 10^{-3} while the valid ones have an average K–S test P value above 0.1 (Fig. 2f). Using 0.01 as the P value cutoff, BiomedParse can achieve an estimated performance of 0.93 precision and 1.00 recall on detecting invalid input (Fig. 2g). BiomedParse substantially outperformed Grounding DINO on invalid input detection (Fig. 2h,i). This enables BiomedParse to perform recognition by enumerating candidate object types in the ontology, skipping invalid text prompts and generating segmentation masks for valid object labels.

Accurate segmentation of irregular-shaped objects

In the previous section, we show that BiomedParse outperformed bounding-box-based methods in general. Additionally, as BiomedParse learns semantic representation for individual object types, we hypothesize that its superiority over previous methods will be even more pronounced in segmenting irregular-shaped objects. To verify this, we show the aggregate attention map of each object type learned by BiomedParse on test images unseen during training and observed that they faithfully reflect object shapes, including many irregular-shaped objects (Fig. 3a). Next, we define three metrics to assess the regularity of an object, including convex ratio (the ratio of the object size to the tightest convex size), box ratio (the ratio of the object size to the tightest rectangle size) and rotational inertia (the difficulty in changing the rotational velocity) (Methods). We found that the improvements of BiomedParse over SAM and MedSAM are strongly correlated with these metrics (average correlation 0.870), indicating that our method has a larger improvement on irregular-shaped objects (Fig. 3b–d and Extended Data Fig. 7). We also found that BiomedParse achieves larger improvement on objects with smaller size (Supplementary Fig. 8). Figure 3e illustrates a few examples comparing BiomedParse and MedSAM on detecting irregular-shaped objects. Furthermore, we show that BiomedParseData has higher average object irregularity than the datasets used by MedSAM (Fig. 3f,g and Supplementary Fig. 9), and the improvement of BiomedParse is also larger on BiomedParseData (Fig. 3h), highlighting the benefit from joint learning of object semantics in detecting the more challenging irregular-shaped objects.

Object recognition using the segmentation ontology

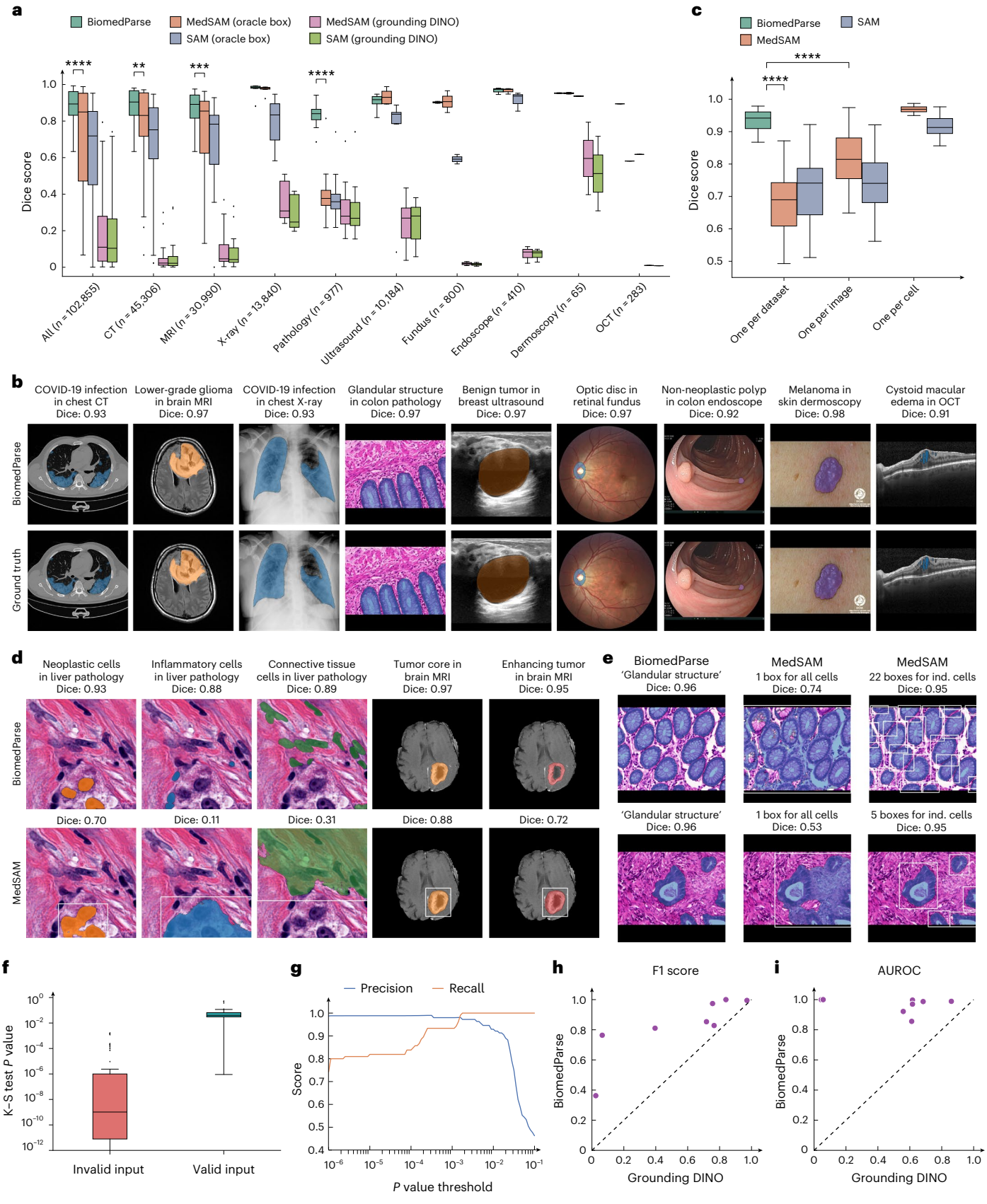
In our final analysis, we explore BiomedParse's capacity for object recognition, which aims to simultaneously segment and label every object within an image. Provided with an image, along with its modality and anatomical site, BiomedParse iteratively performs detection and segmentation for all candidate object types within the ontology of that modality and anatomical site, and the segmented masks are

Fig. 2 | Comparison on large-scale biomedical image segmentation datasets. **a**, Box plot comparing the Dice score between our method and competing methods on 102,855 test instances (image–mask–label triples) across nine modalities. MedSAM and SAM require bounding box as input. We consider two settings: oracle bounding box (minimum bounding box covering the gold mask); bounding boxes generated from the text prompt by Grounding DINO, a state-of-the-art text-based grounding model. Each modality category contains multiple object types. Each object type was aggregated as the instance median to be shown in the plot. n in the plot denotes the number of test instances in the corresponding modality. Significance levels at which BiomedParse outperforms the best-competing method, with two-sided paired t-test are $^{**}P < 1 \times 10^{-2}$; $^{***}P < 1 \times 10^{-3}$; and $^{****}P < 1 \times 10^{-4}$. Exact P values for the comparison between BiomedParse and MedSAM with oracle box prompt are: $P < 1.86 \times 10^{-12}$ for All; $P < 2.49 \times 10^{-3}$ for CT; $P < 3.33 \times 10^{-4}$ for MRI; and $P < 3.30 \times 10^{-16}$ for Pathology. **b**, Nine examples comparing the segmentation results by BiomedParse and the ground truth, using just the text prompt at the top. **c**, Box plot comparing the Dice score between our method and competing methods on a cell segmentation test set with $n = 42$ images. BiomedParse requires only a single user operation (the text prompt 'Glandular structure in colon pathology'). By contrast, to get competitive results, MedSAM and SAM require 430 operations (one bounding box per an individual cell). Significance levels at which BiomedParse outperforms MedSAM, with one-sided paired t-test are $^{**}P < 1 \times 10^{-2}$; $^{***}P < 1 \times 10^{-3}$; and $^{****}P < 1 \times 10^{-4}$. Exact P values are: $P < 1.74 \times 10^{-13}$ for one per dataset and $P < 1.71 \times 10^{-7}$ for one per image. **d**, Five examples contrasting the segmentation results by BiomedParse and MedSAM, along with text prompts used by BiomedParse and bounding boxes used by MedSAM. **e**, Comparison between BiomedParse and MedSAM on a benign tumor image (top) and a malignant tumor image (bottom). The improvement of BiomedParse over MedSAM is even more pronounced on abnormal cells with irregular shapes. **f**, Box plot comparing the two-sided K–S test P values between valid text prompt and invalid text prompt. BiomedParse learns to reject invalid text prompts describing object types not present in the image (small P value). We evaluated a total of 4,887 invalid prompts and 22,355 valid prompts. **g**, Plot showing the precision and recall of our method on detecting invalid text prompts across different K–S test P value cutoff. **h**,**i**, Scatter-plots comparing the area under the receiver operating characteristic curve (AUROC) (**h**) and F1 (**i**) between BiomedParse and Grounding DINO on detecting invalid descriptions. In all box plots, each box shows the quartiles of the distribution, with center as the median, minimum as the first quartile, and maximum as the third quartile. The whiskers extend to the farthest data point that lies within 2 × interquartile range (IQR) from the nearest quartile. Data points that lie outside the whiskers are shown as fliers.

https://doi.org/10.1038/s41592-024-02499-w

aggregated to ensure spatial cohesion among adjacent pixels (Methods). This enables BiomedParse to accurately conduct object recognition, as evidenced in Fig. 4a, where objects are accurately identified and segmented with an average Dice score of 0.94.

Grounding DINO[19] is a state-of-the-art general-domain object recognition system but it does not perform segmentation, which makes Grounding DINO and BiomedParse not directly comparable. We circumvent this by casting the object recognition task as a binary



classification problem: given an input image and a candidate object type, the model determines whether the image contains at least one object of the given type. In this classification formulation, we observed that BiomedParse substantially outperformed Grounding DINO with a 25.0%, 87.9%, 74.5% improvement on precision, recall and F1, respectively (Fig. 4b–d). The improvement over Grounding DINO is even larger when more objects are present in the image (Fig. 4e).

Next, we evaluated the performance of BiomedParse on end-to-end object recognition using weighted average Dice score. Compared to MedSAM and SAM using Grounding DINO for recognition and bounding box generation, BiomedParse outperformed them by a large margin (Fig. 4f and Supplementary Fig. 10). Similar to our observation on object identification, the improvement over comparison approaches is even larger when more objects are present in the image (Fig. 4g). These results indicate BiomedParse's ability to identify all objects in an image, offering an effective tool for holistic image analysis.

Finally, we evaluated BiomedParse on real-world data from the Providence Health System (Fig. 5). We performed object recognition by asking BiomedParse to identify and segment all relevant cells in the pathology slides. We found that the annotations by BiomedParse correctly identified regions of immune cells and cancer cells, attaining high consistency with the pathologist annotations. While pathologists tend to focus on a specific region of the cell type and provide coarse-grained annotations, BiomedParse can precisely label all relevant cells as specified in the ontology, indicating the potential for BiomedParse to help alleviate clinician burdens in real-world clinical applications.

Discussion

We have presented BiomedParse, a biomedical foundation model for image analysis based on image parsing, and a large-scale image parsing dataset BiomedParseData, containing 3.4 million image–mask–label triples and 6.8 million image–mask–description triples. In contrast to existing biomedical foundational models that require users to provide a tight bounding box for each object to segment, BiomedParse is bounding-box-free and can perform holistic image analysis with segmentation, detection and recognition all at once. We conducted a large-scale evaluation on 102,855 test image–mask–label triples across nine modalities. BiomedParse attained new state-of-the-art results, substantially outperforming previous best methods such as MedSAM and SAM, even when they were equipped with an oracle bounding box as the input. The improvement is even larger when the objects have irregular shapes or when an image contains a large number of objects. We also validated the accuracy and scalability of BiomedParse on previously unseen real-world data from the Providence Health System. While BiomedParse has a comparable performance to the state-of-the-art specialized model nnU-Net on most imaging modalities (Supplementary Fig. 6), BiomedParse achieves such promising performance by only using one universal model as opposed to 106 individually trained nnU-Nets models. Collectively, BiomedParse offers an accurate, scalable and robust biomedical image analysis tool that can be broadly applied to various modalities and applications, paving the way for image-based biomedical discovery.

The image analysis field has witnessed rapid development in the past decade. Since its inception in 2015, the U-Net architecture has revolutionized the field of automatic pixel-wise prediction through supervised training[31,32]. This groundbreaking work laid the foundation for a diverse array of network structures, ranging from advanced convolution-network designs to vision-transformer models[27,28,33–47]. Recent advances in image detection and recognition, such as developments in object detection frameworks like Faster R-CNN[48] and YOLOv4 (ref. 49), have enhanced capabilities in identifying and localizing anatomical features with high precision. The introduction of SAM marked a milestone by demonstrating the model's ability to generalize segmentation to previously unseen classes, utilizing visual prompts such as points and bounding boxes as guides[10].

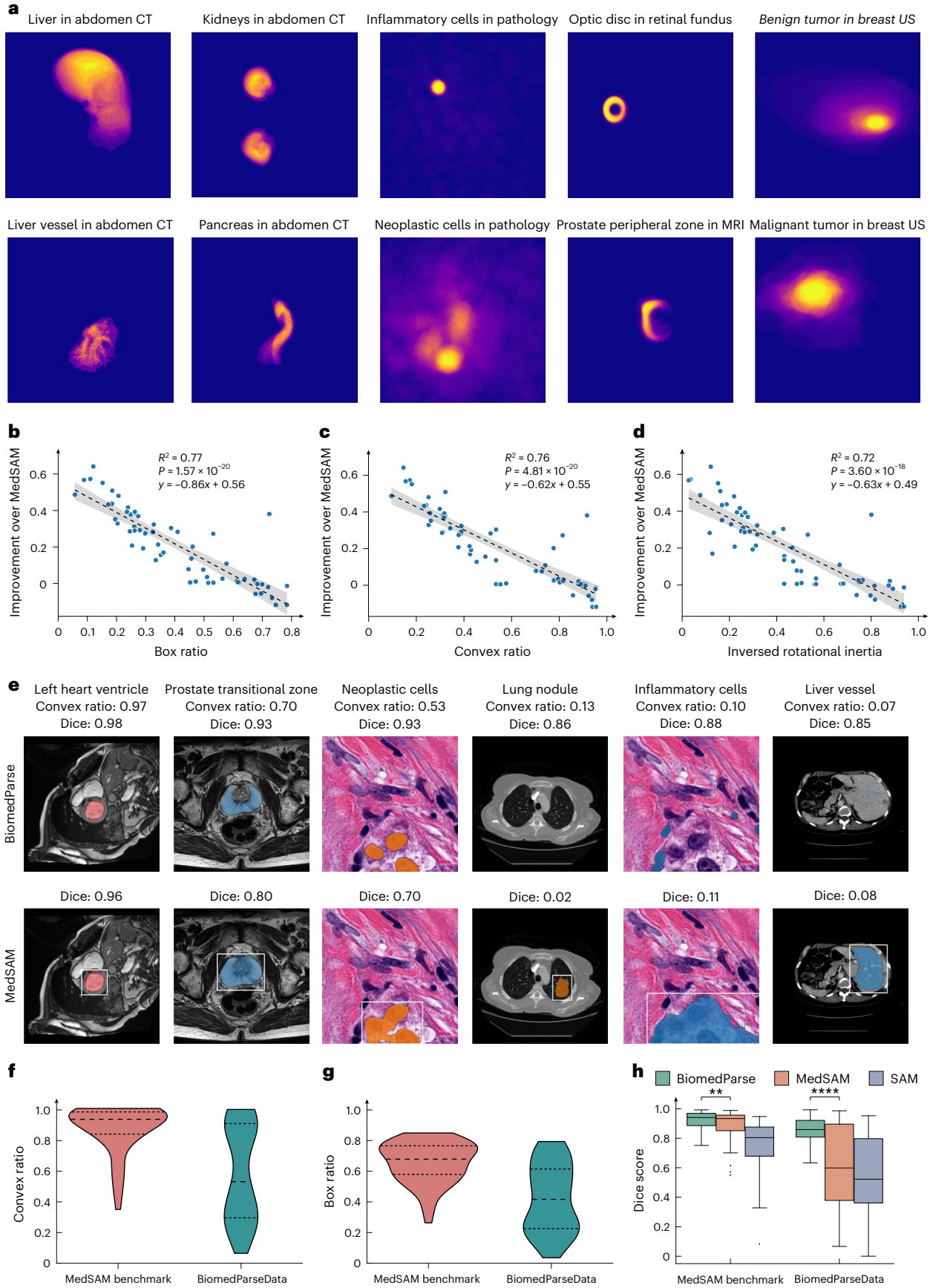
Despite the proliferation of advances in the general domain, research on adapting them for large-scale biomedical image analysis across a wide range of organ or tissue classes remains relatively sparse[50]. MedSAM is a notable exception by adapting SAM to the medical realm through continued training on a large number of biomedical segmentation datasets, establishing the state of art in biomedical image analysis; however, like SAM, MedSAM focuses on segmentation alone, thus ignoring valuable semantic information from related tasks of detection and recognition. Consequently, both SAM and MedSAM require users to provide labor-intensive input such as the tight bounding box for each object to segment, which is hard to scale and very challenging for objects with irregular shapes[11]. As a result, recent works have exploited other types of user operations to replace bounding box[51] or segmentation without bounding boxes[52,53] as alternatives to bounding box-based approaches.

We propose BiomedParse to overcome these challenges due to the bounding boxes. By joint learning across segmentation, detection and recognition in the unifying framework of image parsing, and by using GPT-4 to harmonize noisy object descriptions, BiomedParse was able to acquire new capabilities such as identifying and segmenting objects of interest using a text prompt alone, as well as recognizing all objects in an image by leveraging the segmentation ontology. This represents an important step toward scaling holistic image analysis in biomedicine and real-world clinical applications. If the user has a specific target object type in mind, BiomedParse can perform object detection and segmentation based on the text prompt alone, which specifies the desired object type (Fig. 2b). Alternatively, BiomedParse can be used to identify all available object types without requiring any user text prompt. Behind the scenes, BiomedParse enumerates all possible object types to perform object detection and segmentation simultaneously.

A particularly exciting area for biomedical image analysis is the application in cellular images such as hematoxylin and eosin staining and multiplexed immunofluorescence (MxIF) imaging. This could help elucidate the size, shape, texture and spatial relationships of individual cells, with potential ramifications in emerging applications such as modeling

Fig. 3 | Evaluation on detecting irregular-shaped objects. a, Attention maps of text prompts for irregular-shaped objects, suggesting that BiomedParse learns rather faithful representation of their typical shapes. US, ultrasound. **b**–**d**, Scatter-plots comparing the improvement in Dice score for BiomedParse over MedSAM with shape regularity in terms of convex ratio (**b**), box ratio (**c**) and inversed rotational inertia (**d**). A smaller number in the *x* axis means higher irregularity on average. Each dot represents an object type. We show the regression plot with the 95% confidence interval as the error bands. The *P* values show the two-sided Wald test results. **e**, Six examples contrasting BiomedParse and MedSAM on detecting irregular-shaped objects. Plots are ordered from the least irregular one (left) to the most irregular one (right). **f**,**g** Comparison between BiomedParseData and the benchmark dataset used by MedSAM in terms of convex ratio (**f**) and box ratio (**g**). BiomedParseData is a more faithful representation of real-world challenges in terms of irregular-shaped objects. **h**, Box plots comparing BiomedParse and competing approaches on BiomedParseData and the benchmark dataset used by MedSAM. BiomedParse has a larger improvement on BiomedParseData, which contains more diverse images and more irregular-shaped objects. The number of object types are as follows: $n = 50$ for MedSAM benchmark and $n = 112$ for BiomedParseData. Significance levels at which BiomedParse outperforms the competing method, with a two-sided paired *t*-test are $^{**}P < 1 \times 10^{-2}$ and $^{****}P < 1 \times 10^{-4}$. Exact *P* values were $P < 2.98 \times 10^{-3}$ for MedSAM benchmark and $P < 1.86 \times 10^{-12}$ for BiomedParseData. Each box shows the quartiles of the distribution, with center as the median, minimum as the first quartile, and maximum as the third quartile. The whiskers extend to the farthest data point that lies within 2 × IQR from the nearest quartile. Data points that lie outside the whiskers are shown as fliers.

https://doi.org/10.1038/s41592-024-02499-w



https://doi.org/10.1038/s41592-024-02499-w

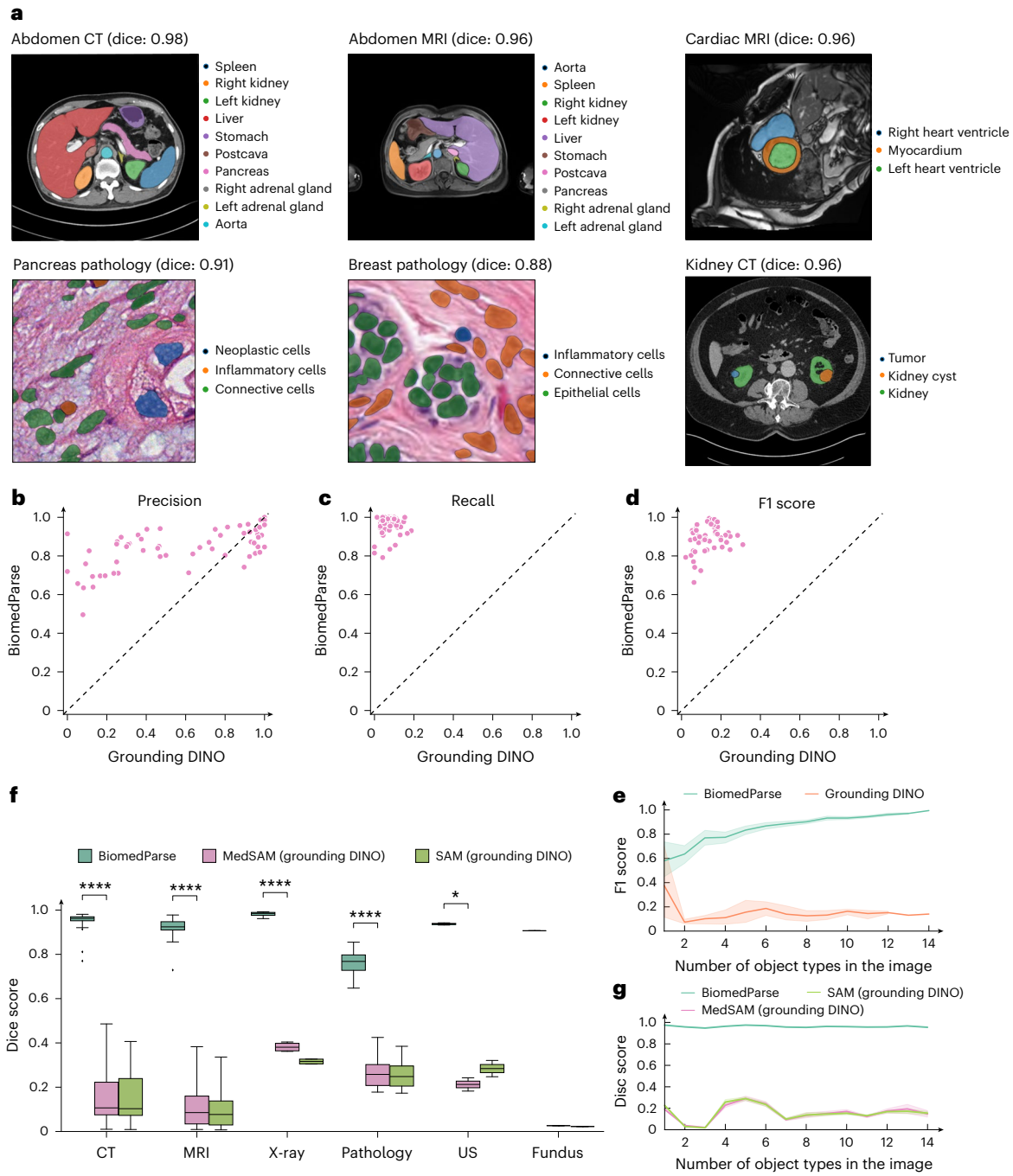


Fig. 4 | Evaluation on object recognition. a, Six examples showing the results of object recognition by our method. Object recognition identifies and segments all objects in an image without requiring any user-provided input prompt. **b**–**d**, Scatter-plots comparing the F1 (**b**), Precision (**c**) and Recall (**d**) scores between BiomedParse and Grounding DINO on identifying objects presented in the image. **e**, Comparison between BiomedParse and Grounding DINO on object identification in terms of median F1 score across different numbers of objects in the image. We show the line plot with the 95% confidence interval as the error bands. **f**, Box plot comparing BiomedParse and MedSAM/SAM (using bounding boxes generated by Grounding DINO) on end-to-end object recognition (including segmentation) in relation to various modalities. Each box shows the quartiles of the distribution, with center as the median, minimum as the first quartile, and maximum as the third quartile. The whiskers extend to the farthest data point that lies within 2 × IQR from the nearest quartile. Data points that lie outside the whiskers are shown as fliers. Each modality category contains image instances with different sets of objects. Each object set was aggregated as the instance median to be shown in the plot. The number of object sets in each modality were as follows: $n = 66$ for CT, $n = 25$ for MRI, $n = 4$ for X-ray, $n = 20$ for Pathology, $n = 2$ for US and $n = 1$ for Fundus. Significance levels at which BiomedParse outperforms the competing method, with a two-sided paired *t*-test *are* $^{**}P < 1 \times 10^{-2}$; $^{***}P < 1 \times 10^{-3}$; and $^{****}P < 1 \times 10^{-4}$. Exact *P* values for the comparison between BiomedParse and MedSAM were $P < 1.96 \times 10^{-57}$ for CT, $P < 4.16 \times 10^{-22}$ for MRI, $P < 3.43 \times 10^{-6}$ for X-ray, $P < 9.42 \times 10^{-20}$ for Pathology $P < 2.19 \times 10^{-2}$ for US. **g**, Comparison between BiomedParse and MedSAM/SAM (using bounding boxes generated by Grounding DINO) on end-to-end object recognition (including segmentation) in relation to numbers of distinct objects in the image. We show the line plot with the 95% confidence interval as the error bands.

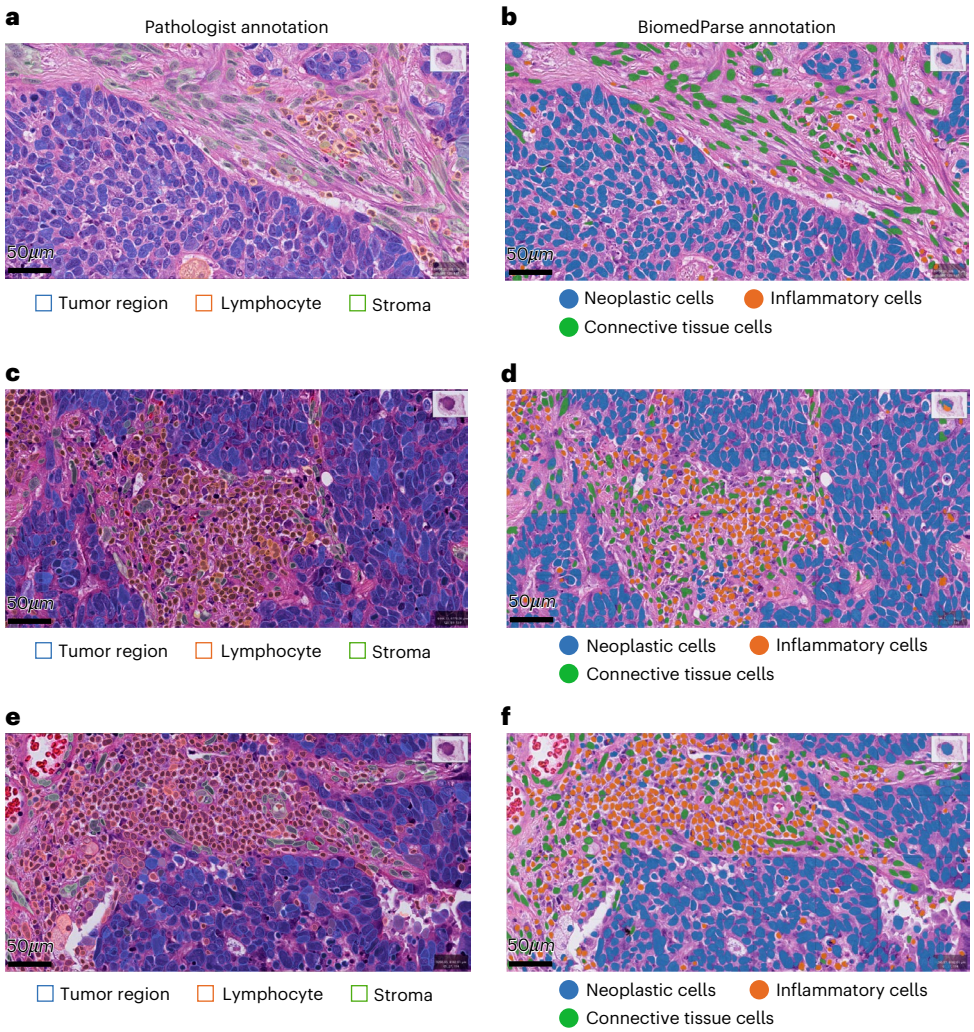


Fig. 5 | Evaluation of BiomedParse on real-world cell segmentation examples. **a**–**f**, De-identified pathology images from the Providence Health System are used to compare pathologist annotations (**a**,**c**,**e**) and annotations from BiomedParse (**b**,**d**,**f**). We show the exact pathologist outputs, including object names (for example, lymphocyte and stroma) and object locations, as well as the exact outputs by BiomedParse. BiomedParse does not need any user-provided text prompt and can identify and segment cells of any types included in the ontology.

tumor microenvironments for precision immunotherapy[54–56]. The standard approaches focus on instance segmentation by assigning unique identifiers to individual cells to facilitate downstream analysis[57–59]. Hover-net represents a notable advancement in addressing the limitations of semantic breadth and cell categorization within segmentation tasks, by incorporating cell classification into the segmentation process[60]; however, traditional methods typically rely on bounding box detection and struggle with diverse cell morphologies and irregular shapes. Recent efforts aim to overcome these challenges by adopting more refined representations and accommodating the multi-resolution nature of biological imaging[61–63]. CellViT is a marquee example that leverages SAM's encoder backbone to improve hierarchical representation, particularly for nucleus segmentation[26]. BiomedParse can contribute to this long line of exciting research work by enabling cell segmentation and identification in one fell swoop and enhancing generalizability through joint training on a diverse range of image modalities and cell types.

While BiomedParse has demonstrated promising potential for unifying biomedical image analysis, growth areas abound. First, although BiomedParse has demonstrated high accuracy (for example, Dice scores) in identifying relevant pixels in an image for a given object type, by default it does not differentiate individual object instances and requires post-processing to separate the instance masks, which is important in some applications such as cell counting. Second, while BiomedParse can already perform image analysis from text prompt alone, it currently does not support interactive dialog with users in a conversational style like GPT-4. To address this, we plan to develop a conversational system that can better tailor to complex user needs. Finally, BiomedParse currently treats non-two-dimensional (2D) modalities such as CT and magnetic resonance imaging (MRI) by reducing them to 2D slices, thus failing to utilize the spatial and temporal information in the original modalities. In future work, we need to extend BiomedParse beyond 2D image slices to facilitate three-dimensional (3D) segmentation, detection and recognition.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-024-02499-w.

References

1. Royer, L. A. The future of bioimage analysis: a dialog between mind and machine. *Nat. Methods* **20**, 951–952 (2023).

2. Li, X., Zhang, Y., Wu, J. & Dai, Q. Challenges and opportunities in bioimage analysis. *Nat. Methods* **20**, 958–961 (2023).
3. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**,181–188 (2024).
4. Liu, Z. et al. OCTCube: a 3D foundation model for optical coherence tomography that improves cross-dataset, cross-disease, cross-device and cross-modality analysis. Preprint at https://www.arxiv.org/abs/2408.11227 (2024).
5. Wang, R. et al. Medical image segmentation using deep learning: a survey. *IET Image Process.* **16**, 1243–1267 (2022).
6. Salpea, N., Tzouveli, P. & Kollias, D. Medical image segmentation: a review of modern architectures. In *European Conference on Computer Vision* 691–708 (Springer, 2022).
7. Ribli, D., Horváth, A., Unger, Z., Pollner, P. & Csabai, I. Detecting and classifying lesions in mammograms with deep learning. *Sci. Rep.* **8**, 4165 (2018).
8. Ma, W., Lu, J. & Wu, H. Cellcano: supervised cell type identification for single cell atac-seq data. *Nat. Commun.* **14**, 1864 (2023).
9. Jiang, H. et al. A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. *Comput. Biol. Med.* **157**, 106726 (2023).
10. Kirillov, A. et al. Segment anything. In *Proc. of the IEEE/CVF International Conference on Computer Vision* 4015–4026 (IEEE, 2023).
11. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
12. Tu, Z., Chen, X., Yuille, A. L. & Zhu, S.-C. Image parsing: Unifying segmentation, detection, and recognition. *Int. J. Comput. Vis.* **63**, 113–140 (2005).
13. Tighe, J. & Lazebnik, S. Superparsing: scalable nonparametric image parsing with superpixels. *Int. J. Comput. Vis.* **101**, 329–349 (2013).
14. Zhou, S. K. *Medical Image Recognition, Segmentation and Parsing: Machine Learning and Multiple Object Approaches* (Academic Press, 2015).
15. Gamper, J. et al. PanNuke dataset extension, insights and baselines. Preprint at https://arxiv.org/abs/2003.10778 (2020).
16. Ji, Y. et al. Amos: a large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Adv. Neural Inf. Process. Syst.* **35**, 36722–36732 (2022).
17. Bernard, O. et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans. Med. Imaging* **37**, 2514–2525 (2018).
18. Lee, H. H. et al. Foundation models for biomedical image segmentation: a survey. Preprint at https://arxiv.org/abs/2401.07654 (2024).
19. Liu, S. et al. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. Preprint at https://arxiv.org/abs/2303.05499 (2023).
20. Zou, X. et al. Segment everything everywhere all at once. In *Proc. 37th Int. Conference on Neural Information Processing Systems* 19769–19782 (Curran Associates, 2024).
21. Yang, J., Li, C., Dai, X. & Gao, J. Focal modulation networks. *Adv. Neural Inf. Process. Syst.* **35**, 4203–4217 (2022).
22. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 1–23 (2021).
23. Sirinukunwattana, K., Snead, D. R. J. & Rajpoot, N. M. A stochastic polygons model for glandular structures in colon histology images. *IEEE Trans. Med. Imaging* **34**, 2366–2378 (2015).
24. Du, Y., Bai, F., Huang, T. & Zhao, B. Segvol: universal and interactive volumetric medical image segmentation. Preprint at https://arxiv.org/abs/2311.13385 (2023).
25. Zhao, Z. et al. One model to rule them all: towards universal segmentation for medical images with text prompts. Preprint at https://arxiv.org/abs/2312.17183 (2023).
26. Hörst, F. et al. Cellvit: vision transformers for precise cell segmentation and classification. *Med. Image Anal.* **94**, 103143 (2024).
27. Hatamizadeh, A. et al. Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In *Int. MICCAI Brain Lesion Workshop* 272–284 (Springer, 2022).
28. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
29. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal.Mach. Intell.* **40**, 834–848 (2017).
30. Butoi, V. I. et al. Universeg: universal medical image segmentation. In *Proc. IEEE/CVF International Conference on Computer Vision* 21438–21451 (ICCV, 2023).
31. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th Int. Conf. Proc. Part III* 234–241 (Springer, 2015).
32. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Int. Conf. Medical Image Computing and Computer-assisted Intervention* 424–432 (Springer, 2016).
33. Milletari, F., Navab, N. & Ahmadi, S.-A. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In *2016 4th Int. Conf. 3D vision (3DV)* 565–571 (IEEE, 2016).
34. Li, X. et al. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* **37**, 2663–2674 (2018).
35. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. UNet++: redesigning Skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **39**, 1856–1867 (2019).
36. Myronenko, A. 3D MRI brain tumor segmentation using autoencoder regularization. In *Int. MICCAI Brain Lesion Workshop* 311–320 (Springer, 2018).
37. Lee, H. H., Bao, S., Huo, Y. & Landman, B. A. 3D UX-Net: a large kernel volumetric ConvNet modernizing hierarchical transformer for medical image segmentation. In *The Eleventh International Conference on Learning Representations* https://iclr.cc/media/iclr-2023/Slides/11340.pdf (ICLR, 2023).
38. Lee, H. H. et al. Scaling up 3D kernels with bayesian frequency re-parameterization for medical image segmentation. In *Int. Conf. Medical Image Computing and Computer-Assisted Intervention* 632–641 (Springer, 2023).
39. Chen, J. et al. TransUNet: transformers make strong encoders for medical image segmentation. Preprint at https://arxiv.org/abs/2102.04306 (2021).
40. Xu, G., Zhang, X., He, X. & Wu, X. LeViT-UNet: make faster encoders with transformer for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)* 42–53 (Springer, 2023).
41. Xie, Y., Zhang, J., Shen, C. & Xia, Y. Cotr: efficiently bridging CNN and transformer for 3D medical image segmentation. In *Int. Conf. Medical Image Computing And Computer-assisted Intervention* 171–180 (Springer, 2021).
42. Wang, W. et al. TransBTS: multimodal brain tumor segmentation using transformer. In *Int. Conf. Medical Image Computing and Computer-Assisted Intervention* 109–119 (Springer, 2021).

43. Hatamizadeh, A. et al. UNETR: transformers for 3D medical image segmentation. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision* 574–584 (2022).
44. Zhou, H.-Y. et al. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Trans. Image Process.* **32**, 4036–4045 (2023).
45. Cao, H. et al. Swin-UNet: UNet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision* 205–218 (Springer, 2022).
46. Zhang, S. et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. Preprint at https://arxiv.org/abs/2303.00915 (2023).
47. Chaves, J. M. Z. et al. Training small multimodal models to bridge biomedical competency gap: a case study in radiology imaging. Preprint at https://arxiv.org/html/2403.08002v2 (2024).
48. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intel.* https://doi.org/10.1109/TPAMI.2016.2577031 (2017).
49. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. Yolov4: optimal speed and accuracy of object detection. Preprint at https://arxiv.org/abs/2004.10934 (2020).
50. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
51. Wong, H. E., Rakic, M., Guttag, J. & Dalca, A. V. Scribbleprompt: fast and flexible interactive segmentation for any medical image. Preprint at https://arxiv.org/html/2312.07381v2 (2024).
52. Shaharabany, T., Dahan, A., Giryes, R. & Wolf, L. AutoSAM: adapting SAM to medical images by overloading the prompt encoder. Preprint at https://arxiv.org/abs/2306.06370 (2023).
53. Lei, W., Wei, X., Zhang, X., Li, K. & Zhang, S. MedLSAM: localize and segment anything model for 3D medical images. Preprint at https://arxiv.org/abs/2306.14752 (2023).
54. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
55. Greenwald, N. F. et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**, 555–565 (2022).
56. Ma, J. & Wang, B. Towards foundation models of biological image segmentation. *Nat. Methods* **20**, 953–955 (2023).
57. Girshick, R. Fast r-cnn. In *Proc. IEEE Int. Conf. on Computer Vision* 1440–1448 (IEEE, 2015).
58. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. In *Proc. IEEE Int. Conf. On Computer Vision* 2961–2969 (IEEE, 2017).
59. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st Int. Conf. Proc. Part II* 265–273 (Springer, 2018).
60. Graham, S. et al. Hover-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).
61. Yang, H. et al. CircleNet: anchor-free glomerulus detection with circle representation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd Int. Conf. Proc. Part IV* 35–44 (Springer, 2020).
62. Nguyen, E. H. et al. CircleSnake: instance segmentation with circle representation. In *Int. Workshop on Machine Learning in Medical Imaging* 298–306 (Springer, 2022).
63. Ilyas, T. et al. Tsfd-net: tissue specific feature distillation network for nuclei segmentation and classification. *Neural Netw.* **151**, 1–15 (2022).

Methods

Details of BiomedParseData

We created a large-scale biomedical image parsing dataset called BiomedParseData, where each image is associated with a collection of objects. Each object is annotated with the segmentation mask and a canonical semantic label specifying the object type from a biomedical object ontology. Additionally, each semantic label comes with a set of synonymous textual descriptions for model training. BiomedParseData was created by synthesizing 45 publicly available biomedical segmentation datasets across nine imaging modalities, comprising 1.1 million images, 3.4 million image–mask–label triples and 6.8 million image–mask–description triples (Fig. 1b and Supplementary Table 1). To ensure the quality of BiomedParseData, we imposed stringent inclusion criteria: each image had to be manually or semi-manually segmented at the pixel level and a name was available for each segmented object from the dataset description. For 3D imaging modalities such as CT and MRI, we pre-processed each volume into in-plane 2D slices to be consistent with other modalities.

For model training and evaluation, we randomly split each original dataset into 80% training and 20% testing. Slices from each 3D volume always appear in the same split to prevent information leakage.

To harmonize natural language variations in noisy object descriptions, we use GPT-4 to assist the creation of a three-layer biomedical object ontology (Fig. 1a). The base layer comprises three broad semantic categories: organ, abnormality and histology. The next layer comprises 15 meta-object types (for example, heart in organ and tumor in abnormality). The most fine-grained layer comprises 82 object types, such as left heart ventricle and enhancing tumor. Specifically, we first used GPT-4 to generate a preliminary hierarchical structure for biomedical image analysis and propose candidate names for individual object types, drawing from a wide range of tasks and textual descriptions across the source datasets. We then manually reviewed these candidates and mapped them to standardized Observational Health Data Sciences and Informatics (OHDSI) vocabularies using Athena[64]. Most of these candidates are mapped to 15 meta-object types by searching in the OHDSI vocabulary. For some of them that cannot be mapped to the meta-object types or the meta-object types do not exist in the OHDSI vocabulary, we asked GPT-4 to suggest the appropriate meta-object type names and do the mapping. We introduce 'other' as a catch-all category. For future expansion, we expect that the first two layers are relatively stable, while our framework can easily incorporate new object types in the fine-grained layers.

To enhance the robustness of BiomedParse in handling diverse text prompts, we also used GPT-4 to generate synonymous textual descriptions for each semantic label, following other recent efforts in using GPT-4 for synthetic data generation[65–67]. Specifically, we adopted a templatic normalization for each dataset by formulating the unifying image analysis task as identifying '[OBJECT TYPE] in [ANATOMIC SITE] [MODALITY]', such as 'enhancing tumor in brain MRI' (Extended Data Fig. 1). We then introduced linguistic diversity into these descriptions by using GPT-4 to generate variations in professional language (Supplementary Fig. 1), as well as introducing synonymous variations for each component (Supplementary Fig. 2). We manually checked all the templates that we used to prompt GPT-4 for variations to avoid incorrectness and hallucinations. We define incorrectness and hallucination as (1) not mentioning the target object; (2) only describing the image; (3) referring to another target; and (4) describing another image modality. We found that the descriptions provided by GPT-4 are generally correct and only less than 10% templates were removed from the initial prompts. For training, the number of prompts depends on the object type, with a minimum prompts of 1, an average prompts of 8.28, a median prompts of 7 and a maximum prompt of 36. We randomly sampled one prompt for training. For inference, we only used one prompt for each data point and used the original description as the prompt. We compared the performance between varying the text prompt at the inference stage and using a fixed one based on the original description and did not observe a statistically different performance (Supplementary Fig. 11). In each training epoch, we randomly sampled a description for each image–mask pair, enabling BiomedParse to understand diverse text prompts.

Details of BiomedParse

Existing image analysis methods often focus on segmentation alone. They typically expect spatial input prompts such as bounding box or scribble for the object to segment and focus on learning spatial embedding such as bounding box coordinates[10,11,51].

In contrast, BiomedParse follows SEEM[20] and focuses on learning text prompts. Specifically, BiomedParse adopts a modular design, comprising an image encoder, a text encoder, a mask decoder and a meta-object classifier (Fig. 1c). We initialized the model from SEEM, with each module described in detail below.

The input to BiomedParse is an image and a text prompt, which are passed along to the image and text encoders, respectively. The text prompt specifies the object type for segmentation and detection in the image. The image encoder processes the high-resolution image and outputs downsampled embeddings. We provide a flexible choice of backbone architectures with Focal[21] and SAM-ViT[10]. The text encoder processes the user-provided prompt and generates language embeddings. We provide options to use the pretrained biomedical language model PubMedBERT[22] or training a transformer from scratch. The base version of BiomedParse adopts Focal as an image encoder and the text encoder transformer fully trained on BiomedParseData.

The mask decoder outputs a segmentation mask that has the same size as the original image, with a probability between 0 and 1 for each pixel, indicating how likely the pixel belongs to the designated object in the text prompt. The meta-object classifier includes input from the image and text prompt and output object semantics. We follow SEEM[20] and X-Decoder[68] to build the segmentation decoder head. The decoder is a transformer that cross-attend the image and text embeddings and gradually upsample the image features back to high-resolution pixels. At the last layer, the attention dot product on the pixel embeddings delivers the segmentation mask.

Details of model training

The training of BiomedParse is around segmentation with grounding text. Therefore, during training time the following linear combination of losses is minimized:

$$\mathcal{L} = a\mathcal{L}_{\text{c\_CE\_text}} + b\mathcal{L}_{\text{m\_BCE\_text}} + c\mathcal{L}_{\text{m\_Dice\_text}}, \tag{1}$$

where c stands for meta-concept classification with cross-entropy loss (CE), m stands for mask prediction with binary cross-entropy and Dice loss. The formula for the losses are as follows:

$$\mathcal{L}_{\text{c\_CE\_text}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \tag{2}$$

$$\mathcal{L}_{\text{m\_BCE\_text}} = -\frac{1}{\mathcal{P}} \sum_{p\in\mathcal{P}} \left(m_p \log(\hat{m}_p) + (1-m_p)\log(1-\hat{m}_p)\right), \tag{3}$$

$$\mathcal{L}_{\text{m\_Dice\_text}} = 1 - \frac{2\sum_{p\in\mathcal{P}} m_p \hat{m}_p}{\sum_{p\in\mathcal{P}} m_p + \sum_{p\in\mathcal{P}} \hat{m}_p}, \tag{4}$$

where y is the one-hot vector of true meta-concept over $c = 1, \cdots, C$ and $\hat{y}$ is the predicted meta-concept probability distribution. m_p is the ground-truth binary mask for pixel $p \in \mathcal{P}$ and $\hat{m}_p$ is the predicted pixel probability. We follow SEEM[20] and append the visual sampler loss and

other auxiliary losses during training to enable interactive spatial refinement, which we refer to the original paper for details. For BiomedParse training, we assign equal weights for the three losses.

We initialized BiomedParse from the pretrained SEEM model. As a result, we follow the exact hyperparameter setting in the SEEM paper to perform continue training for text prompt-based segmentation. Specifically, we fix learning rate of 10^{-5} and train for 20 epochs. To train BiomedParse, we used 16 NVIDIA A100-SXM4-40GB GPUs for a duration of 58 h. We performed inference evaluation with four NVIDIA RTX A6000 GPUs. The inference time with a single NVIDIA RTX A6000 GPU is 0.17 s per data point. The minimum hardware needed for performing the inference is one V100 GPU with 16 GB memory. The post-processing time in the object recognition task are 0.11 s and 0.07 s on average for target selection and mask aggregation stages, respectively.

Mixed dataset training. To enable flexible incorporation of multiple datasets, we perform random mixing at a batch level. We denote each dataset of a modality as $\mathcal{D}_m$ for $m = 1, \cdots, M$. The creation of each batch follows the procedure below:

- In each iteration i, we aggregate a batch from K mini-batches $b_1^i, \cdots, b_K^i$. For each mini-batch b_k^i, we randomly select dataset $\mathcal{D}_m$ for $m = 1, \cdots, M$ with probability p_m and sample the mini-batch without replacement.
- Concatenate all mini-batches $B^i = [b_1^i, \cdots, b_K^i]$.
- Perform training step with batch B^i.

We can flexibly control the training data distribution from all the datasets with the sampling probability p_m. As the size difference of the datasets can be large, we define a parametric probability distribution

$$p_m = \frac{|\mathcal{D}_m|^\lambda}{\sum_{m'=1}^{M} |\mathcal{D}_{m'}|^\lambda}, \quad 0 \le \lambda \le 1. \tag{5}$$

When $\lambda = 1$, we sample the mini-batch with probability proportional to the size of the datasets, thus each example from any dataset has equal chance to be selected. The downside is that the training will be overwhelmed with the huge datasets, while ignoring the smaller ones which are also important.

On the other extreme, when $\lambda = 0$, each dataset has equal chance to be selected in each iteration. This ensures a good diversity of tasks, but the small datasets which have very few examples will be repeated for a large amount of time, causing overfitting to the training examples. On the other hand, the large datasets will never be exhausted in a limited number of epochs.

As a balance between the two extremes, we set $\lambda = 1/2$ for the first five epochs, then set $\lambda = 1$ for the next ten epochs and $\lambda = 1/2$ again for the last five epochs.

Implementation of competing methods

We compared BiomedParse to the state-of-the-art segmentation models, SAM[10] and MedSAM[11]. We recognize the importance of precise bounding boxes as the model input, so we evaluated competing methods in two settings: (1) employing gold-standard bounding boxes, and (2) utilizing bounding boxes predicted by the state-of-the-art object detection model Grounding DINO[19] to provide bounding box prompts. For the first setting, we followed previous work[11] by deriving bounding boxes from gold-standard masks, ensuring each box tightly encompassed the mask with a uniform margin of 10 pixels. In the second setting, we adhered to the inference pipeline of Grounding DINO where, when presented with multiple bounding box predictions, we selected the one with the highest confidence score. This text-to-box-to-segmentation scheme follows the idea of previous work[69]. In addition to comparing current SAM-based state-of-the-art models, we also evaluated BiomedParse against (1) the established medical segmentation approach nnU-Net[28], an end-to-end U-Net architecture that adapts to various medical imaging modalities using a purely convolutional module and fully supervised learning without prompts, and (2) the general domain segmentation architecture DeeplabV3+, which uses ResNet-101 as the architecture backbone with an Atrous Spatial Pyramid Pooling module for decoding and upsampling bottleneck features with multiple fields of view[70]. To maintain uniformity across comparisons, all input images were resized to 1,024 × 1,024 pixels. We use the same test split of BiomedParseData for evaluation across competing methods, and performance was quantified using the median Dice score on each task. We recognize that the train–test splits are different across the original evaluations of the competing methods, and the BiomedParseData test split could contain examples that were used to train other models. We note that the implementations for MedSAM, SAM and Grounding DINO were used as is for inference purposes without any fine-tuning. As for the task-specific nnU-Net models[28] and the DeepLabV3+ models[70], we trained both network architectures in 2D with one binary segmentation model for each target in each modality, resulting in 95 task-specific models for each method. We adopted the built-in automatic hyperparameter configuration in nnU-Net. For the Deeplabv3 network, we trained all models in 50 epochs with batch size of four and a learning rate of 0.0003 with weight decay of 0.0001 using AdamW optimizer.

For continued training MedSAM and SAM experiments, we provided the entire training dataset that was used to train BiomedParse. MedSAM and SAM are provided with oracle bounding boxes during training and inference. We fixed the SAM and MedSAM backbone respectively and further trained for ten epochs each, resulting in SAM-FT and MedSAM-FT, respectively. When evaluating UniverSeg[30], we provided 16 support images for the model as examples, as shown as the optimal number of support images in the original paper. For CellViT[26], we used the PanNuke[15] dataset as the evaluation datasets, which contains cell segmentations across tissue types. We compared BiomedParse to SegVol[24], SAT[25] and Swin UNETR[27] using CT imaging from the Amos22 (ref. 16) dataset as SegVol is specific to CT, and SAT is specific to CT, MRI and positron emission tomography. Both SegVol and SAT adapted SAM architecture to 3D medical volume and leverage text of anatomical regions as input besides visual prompts of boxes and points. Swin UNETR, built on Swin transformers, is a widely used benchmark for CT segmentation tasks and achieved top performance on BraTS challenge[71]. We used the strongest available model weights and the same text prompts as specified in the respective papers.

Detecting invalid textual description

BiomedParse by design can input any image and text prompt; however, a text prompt may be invalid, specifying an object that does not exist in the given image[66,72]. For example, the request to identify and segment 'left heart ventricle' in a dermoscopy image should be rejected by the model as invalid. It is critical to detect and reject invalid text prompts to pre-empt hallucinations[73].

In principle, the mask decoder should output low pixel probabilities for invalid text prompt; however, given the sheer number of pixels, some might get a relatively high output probability simply by chance, thus leading to erroneous object detection and segmentation results. To address this problem, we observe that while individual pixels might get noisily high probabilities, collectively their distribution would be rather different compared to pixels in valid objects. Consequently, we can estimate the distribution of its pixel probabilities from training data, and then estimate how likely the pixel probabilities in a test image are drawn from the same distribution.

Specifically, after BiomedParse was trained, for each object type, we computed the average object pixel probability for each training image containing objects of the given type, and fit a beta distribution for all these probabilities. At test time, for a given image, we computed

the average object pixel probability for the predicted object segments of the given object type, and compute the P value using a one-sample K–S test[74]. A smaller P value indicates that the predicted object segments are unlikely to be correct. To increase the robustness, in addition to pixel probability, we also consider the RGB values. In particular, for each color channel (R, G and B), we similarly fit a beta distribution from the average value for valid objects in training, and compute the corresponding P value for the predicted object segments in a test image. Overall, we treat these four tests as independent and use their product as the summary P value.

In this way, we can obtain a summary P value for any given pair of text prompt and image. To identify a summary P value threshold for separating valid inputs from invalid ones, we created an invalid dataset by mixing datasets of different modalities. For example, we take a target in a certain modality (for example heart anatomies in MRI), and apply the corresponding text prompt to identify this target in another modality (for example endoscopy) if this target has never appeared in that modality. The text prompts for heart anatomies are now invalid in the endoscopy dataset, providing us valid examples (prompts in the original modality) and invalid examples (prompts in the alternative modality). We plot the distribution for both valid text prompts (for a given image) and invalid ones (Fig. 2f). For comparison against Grounding DINO, we use its confidence score given a text prompt and an image for invalid input detection.

Attention map conditioned on the textual description

To visualize the shape of each segmentation object type, for example 'hepatic vessel in CT', we collected the predicted pixel probabilities for each object type and aggregated probabilities from all images. The pixel-level probability is derived from the top layer attention on the pixel. The attention map, reflecting the shape for a target t, is obtained in a four-step approach. First, we collected all BiomedParse-predicted pixel attention for target t as $\rho_1, \cdots, \rho_n \in [0,1]^{H\times W}$ across n examples in the test set. Second, we initialized shape distribution for target t as $\mathcal{M}_1^t = \rho_1$. Third, for iteration $i = 1, \cdots, n-1$, we computed 2D cross-correlation between ρ_{i+1} and $\mathcal{M}_i^t$ and shifted ρ_{i+1} to be aligned with $\mathcal{M}_i^t$ at highest cross-correlation, and updated the ensemble distribution $M_{i+1}^t = M_i^t + \tilde{\rho}_{i+1}$, where $\tilde{\rho}_{i+1}$ denotes the shifted attention matrix. Finally, the attention map for target t is normalized as M_n^t/n. For 3D segmentation targets such as CT and MRI, we first aggregated the predictions within one volume without shifting and then aligned the volume-aggregated masks using the above method.

Details of experiments on irregular-shaped objects

Medical image segmentation models like MedSAM require a bounding box as input. When the shape of the target is 'irregular', it is hard for the bounding box to precisely define the region of interest. To quantify the 'regularity' of a target mask M, we define the following three metrics: Box Ratio measures the degree to which the target mask is similar to its tight bounding box: $\text{BoxRatio}(M) = \frac{|M|}{|\text{Box}(M)|}$, where $\text{Box}(M)$ is the tight bounding box around mask M and $|\cdot|$ denotes the area measured in number of pixels. Convex Ratio measures how convex the target mask is and is defined as $\text{ConvexRatio}(M) = \frac{|M|}{|\text{ConvexHull}(M)|}$, where $\text{ConvexHull}(M)$ is the convex hull of mask M. Convex hull is defined as the intersection of all convex sets containing a given subset of a Euclidean space. In other words, it is the smallest convex region that covers the shape. Inverse rotational inertia (IRI) measures how spread out the area of the target mask is. To begin with, the rotational inertia (RI) of M relative to its centroid c_M is $\text{RI}(M) = \sum_{x\in M} \| x - c_M \|_2^2$, where x is the coordinate of each pixel in the mask and c_M is the coordinate of the centroid. To standardize the metric to be independent of the total mask area, we take the inverse of the RI and scale by the value of a round-shaped mask with the same area, representing the lowest rotational inertia achievable by any mask with the same area: $\text{IRI}(M) = \frac{|M|^2}{2\pi\cdot\text{RI}(M)}$. Under this definition, any mask has $0 < \text{IRI} \le 1$, with any round-shaped mask having IRI equal to 1.

Details of experiments on object recognition

We built a hierarchical structure putting all supported targets under one modality at one anatomic site. Given any image, for example abdominal CT, we traverse all the available targets $t = 1, \cdots, m$ under the branch that are exclusive to each other, and prompt the BiomedParse model sequentially to get m prediction of mask probabilities $\rho^1, \cdots, \rho^m$. It is possible that the predicted masks can overlap with each other. The challenges then are how to select the right set of targets in the specific image and how to determine the right mask regions for the selected targets to avoid overlapping. We used a two-stage approach for object recognition, including a target selection stage and a mask aggregation stage. In the target selection stage, we first calculate the original mask area for each target t as A^t. Then, we iterate through the pixels. For each pixel (i,j), we rank the targets that have pixel probability $\rho_{ij}^t > 0.5$. The target assigned to pixel (i,j) is $T_{ij} = \text{argmax}\, \rho_{ij}^t$. After this round of pixel assigning, the final area for each target t is $\tilde{A}^t = \sum_{i,j} \mathbf{1}_{T_{ij}=t}$. The targets with final area $\tilde{A}^t > \lambda A^t$ are the selected targets, with λ being the user-specified threshold. In the mask aggregation stage, we discard all unselected target masks completely and then iterate through the pixels again. For each pixel, the most probable target t with $\rho_{ij}^t > 0.5$ is assigned. The pixels with predicted probabilities $\rho_{ij}^t \le 0.5$ for all selected targets are left blank.

For the baseline method using Grounding DINO with SAM and MedSAM, we first prompted Grounding DINO with the set of targets to retrieve a collection of bounding boxes with confidence scores. Then we implemented nonmaximum suppression[75–77] to select a subset of identified targets in the scene, minimizing the overlapping between the targets. To get the segmentation masks for these identified targets, we further prompted SAM and MedSAM with the bounding boxes to retrieve the corresponding predictions.

Data collection and analysis

All source image data were from publicly available datasets. We used Python (v.3.10.12) to curate and preprocess the image data. For the textual description for the objects in the images, we used GPT-4 provided by Azure OpenAI to generate text data. This work used open-source code bases and libraries to analyze the data. We used SEEM (https://github.com/UX-Decoder/Segment-Everything-Everywhere-All-At-Once) for the main model architecture and training of the model on the datasets. We used matplotlib v.3.8.2 and seaborn v.0.11.2 to visualize the data.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

BiomedParseData can be accessed at https://aka.ms/biomedparse-release. The three real-world pathology images, including the annotations by pathologists and BiomedParse, can be accessed at https://aka.ms/biomedparse-release.

Code availability

BiomedParse can be accessed at https://aka.ms/biomedparse-release, including the model weights and relevant source code. We include detailed methods and implementation steps in the Methods to allow for independent replication.

References

64. OHDSI. Athena standardized vocabularies. https://www.ohdsi.org/analytic-tools/athena-standardized-vocabularies/

65. Gu, Y. et al. BiomedJourney: counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. Preprint at https://arxiv.org/abs/2310.10765 (2023).
66. Li, C. et al. Llava-med: training a large language-and-vision assistant for biomedicine in one day. In *37th Conference on Neural Information Processing Systems* https://proceedings.neurips.cc/paper_files/paper/2023/file/5abcdf8ecdcacba028c6662789194572-Paper-Datasets_and_Benchmarks.pdf (NeurIPS, 2024).
67. Gu, Y., Zhang, S., Usuyama, N. et al. Distilling large language models for biomedical knowledge extraction: a case study on adverse drug events. Preprint at https://arxiv.org/abs/2307.06439 (2023).
68. Zou, X. et al. Generalized decoding for pixel, image, and language. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 15116–15127 (IEEE, 2023).
69. Ren, T. et al. Grounded SAM: assembling open-world models for diverse visual tasks. Preprint at https://arxiv.org/abs/2401.14159 (2024).
70. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. European Conference on Computer Vision (ECCV)* 801–818 (2018).
71. Kazerooni, A. F. et al. The brain tumor segmentation (BraTS) challenge 2023: focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). Preprint at https://arxiv.org/abs/2305.17033 (2023).
72. Lee, P., Goldberg, C. & Kohane, I. *The AI Revolution in Medicine: GPT-4 and Beyond* (Pearson, 2023).
73. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
74. Massey Jr, F. J. The Kolmogorov–Smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **46**, 68–78 (1951).
75. Canny, J. A computational approach to edge detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* 679–698 (IEEE, 1986).
76. Viola, P. & Jones, M. Rapid object detection using a boosted cascade of simple features. In *Proc. 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, I–I (IEEE, 2001).
77. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 580–587 (2014).

Acknowledgements

The authors thank the Microsoft Health and Life Sciences Research team and the Microsoft Health Futures team for support and helpful discussions.

Author contributions

T.Z., Y.G., J.Y., N.U., M.W., H.P. and S.W. contributed to the conception and design of the work. T.Z. contributed to the data acquisition and curation of BiomedParseData. T.Z., J.Y., N.U. and M.W. contributed to BiomedParse model training. Y.G., T.Z., H.L., N.U. and S.K. contributed to the evaluation of BiomedParse and baseline models. T.N. and J.G. contributed to the technical discussions. A.C., J.A., C.M., B.P. and C.B. provided clinical inputs to the study. All authors contributed to the drafting and revision of the manuscript.

Competing interests

C.B. is a member of the scientific advisory board and owns stock in PrimeVax and BioAI; is on the scientific board of Lunaphore and SironaDx; has a consultant or advisory relationship with Sanofi, Agilent, Roche and Incendia; contributes to institutional research for Illumina, and is an inventor on US patent applications US20180322632A1 (Image Processing Systems and Methods for Displaying Multiple Images of a Biological Specimen) filed by Ventana Medical Systems, Providence Health and Services Oregon and US20200388033A1 (System and Method for Automatic Labeling of Pathology Images) filed by Providence Health and Services Oregon, Omics Data Automation. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-024-02499-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02499-w.

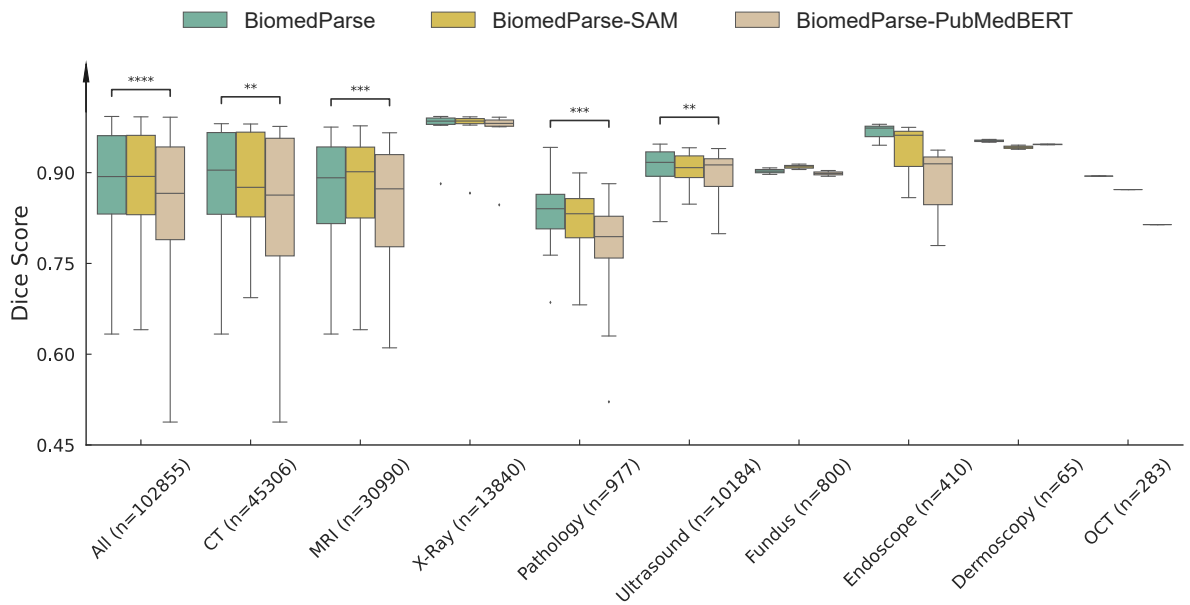
Correspondence and requests for materials should be addressed to Mu Wei, Hoifung Poon or Sheng Wang.

Peer review information *Nature Methods* thanks Stefania Moroianu, Dong Ni, Yichi Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Rita Strack, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

https://doi.org/10.1038/s41592-024-02499-w

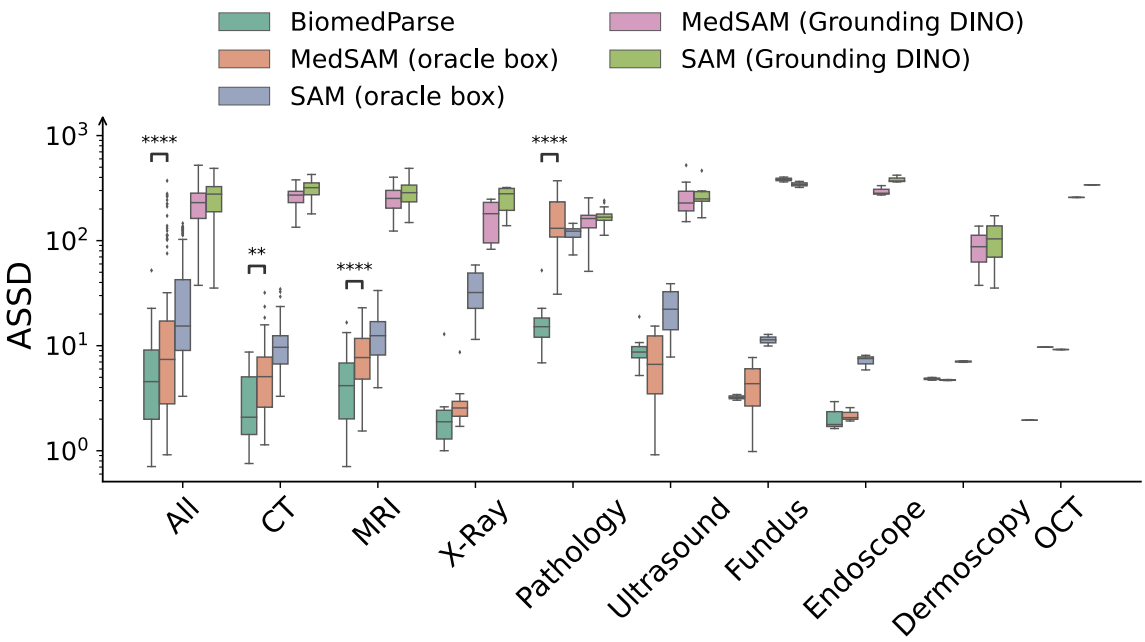
Extended Data Fig. 1 | Number of images in each of the 25 anatomic sites from 9 modalities. One anatomic site could present in multiple modalities.

https://doi.org/10.1038/s41592-024-02499-w



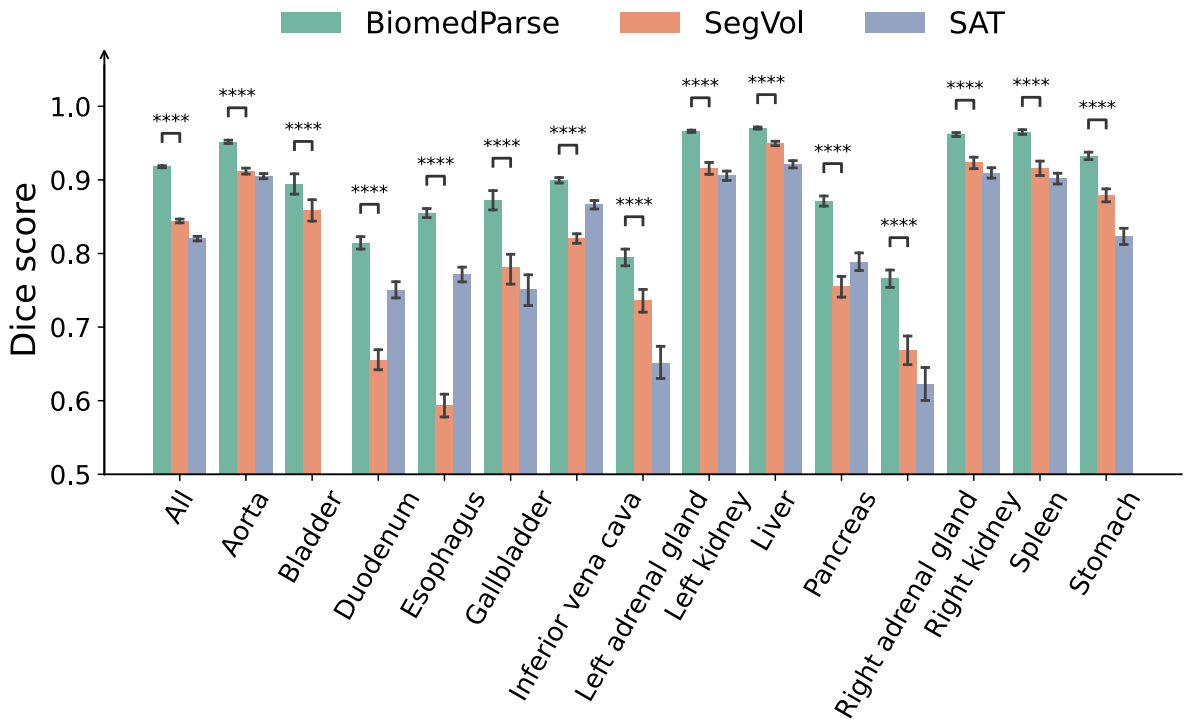
Extended Data Fig. 2 | Ablation studies comparing the performance of BiomedParse and two variants. BiomedParse-SAM stands for using SAM to initialize the image encoder. BiomedParse-PubmedBERT stands for using the frozen PubmedBERT as the text encoder. Each modality category contains multiple object types. Each object type was aggregated as the instance median to be shown in the plot. N in the plot denotes the number of images in the corresponding modality. The numbers of object types in each modality are as follows: N = 112 for All, N = 27 for CT, N = 34 for MRI, N = 12 for X-Ray, N = 24 for Pathology, N = 7 for Ultrasound, N = 2 for Fundus, N = 3 for Endoscope, N = 2 for Dermoscopy, and N = 1 for OCT. Each box shows the quartiles of the distribution, with the center as the median, the minimum as the first quartile, and the maximum as the third quartile. The whiskers extend to the farthest data point that lies within 2 times the inter-quartile range (IQR) from the nearest quartile. Data points that lie outside the whiskers are shown as fliers. *indicates the significance level at which BiomedParse outperforms BiomedParse-PubmedBERT, with two-sided paired t-test p-value $< 1 \times 10^{-2}$ for **, p-value $< 1 \times 10^{-3}$ for ***, p-value $< 1 \times 10^{-4}$ for ****. Exact p-values for the comparison between BiomedParse and BiomedParse-PubMedBERT are as follows: p-value $< 9.52 \times 10^{-10}$ for All, p-value $< 1.67 \times 10^{-3}$ for CT, p-value $< 4.87 \times 10^{-4}$ for MRI, p-value $< 1.98 \times 10^{-4}$ for Pathology, and p-value $< 7.13 \times 10^{-3}$ for Ultrasound.

https://doi.org/10.1038/s41592-024-02499-w



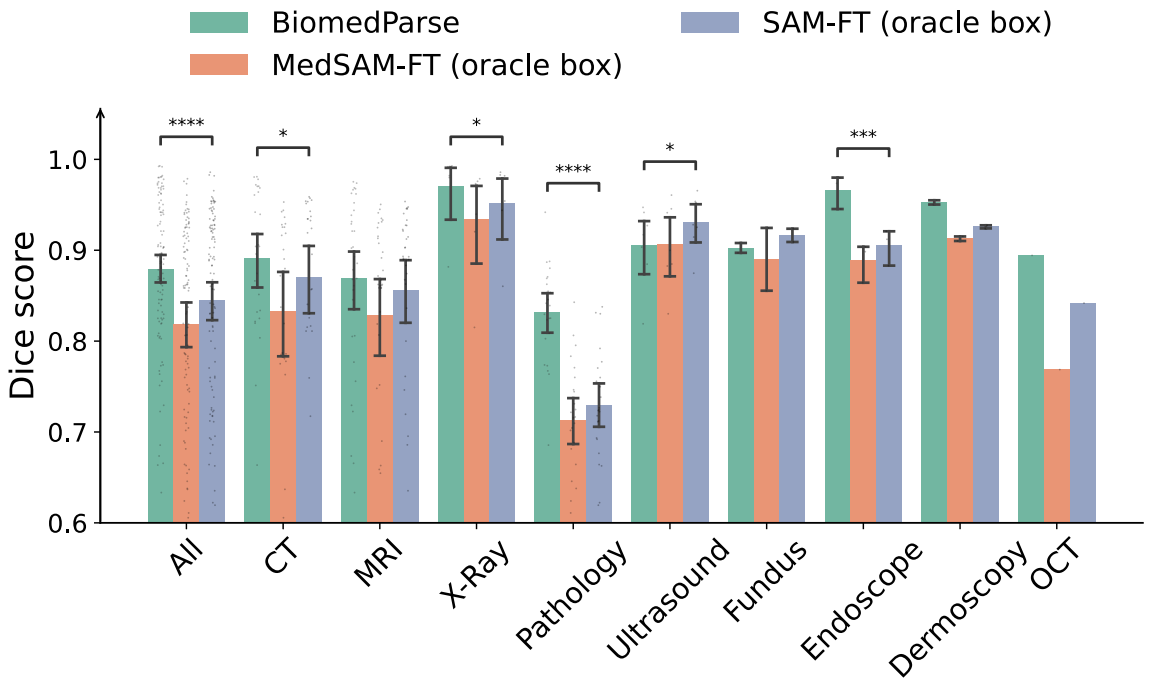
Extended Data Fig. 3 | Evaluating BiomedParse and competing methods in terms of Average Symmetric Surface Distance. Box plot comparing the performance of BiomedParse and competing methods in terms of Average Symmetric Surface Distance (ASSD). Smaller ASSD indicates better segmentation performance. Each box shows the quartiles of the distribution, with center as the median, minimum as the first quartile, and maximum as the third quartile. The whiskers extend to the farthest data point that lies within 2 times the inter-quartile range (IQR) from the nearest quartile. Data points that lie outside the whiskers are shown as fliers. Each modality category contains multiple object types. Each object type was aggregated as the instance median to be shown in the plot. The numbers of object types in each modality are as follows: n = 112 for All, n = 27 for CT, n = 34 for MRI, n = 12 for X-Ray, n = 24 for Pathology, n = 7 for Ultrasound, n = 2 for Fundus, n = 3 for Endoscope, n = 2 for Dermoscopy, and n = 1 for OCT. *indicates the significance level at which BiomedParse outperforms the best-competing method, with two-sided paired t-test p-value $< 1 \times 10^{-2}$ for **, p-value $< 1 \times 10^{-3}$ for ***, p-value $< 1 \times 10^{-4}$ for ****. Exact p-values for the comparison between BiomedParse and MedSAM with oracle box prompt are as follows: p-value $< 3.43 \times 10^{-6}$ for All, p-value $< 2.61 \times 10^{-3}$ for CT, p-value $< 7.73 \times 10^{-5}$ for MRI, and p-value $< 2.94 \times 10^{-8}$ for Pathology.

https://doi.org/10.1038/s41592-024-02499-w



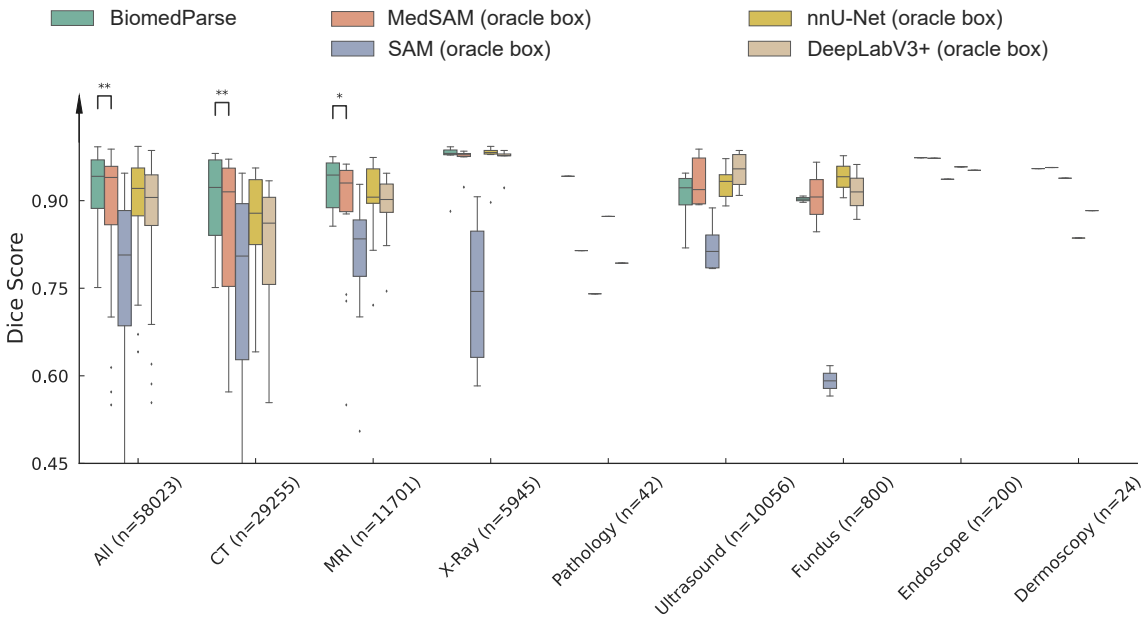
Extended Data Fig. 4 | Comparing BiomedParse with biomedical-specific text prompt segmentation models. Bar plot comparing BiomedParse with biomedical-specific text prompt segmentation models across different organs on CT in terms of Dice score. Each bar shows the mean of the distribution, with error bar indicating the 95% confidence interval. The sample sizes for the target organs are as follows: n = 27,779 for All, n = 4,409 for Aorta, n = 864 for Bladder, n = 1,677 for Duodenum, n = 1,964 for Esophagus, n = 712 for Gallbladder, n = 4,105 for Inferior vena cava, n = 635 for Left adrenal gland, n = 1,776 for Left kidney, n = 4,648 for Liver, n = 1,345 for Pancreas, n = 571 for Right adrenal gland, n = 1,649 for Right kidney, n = 1,587 for Spleen, and n = 1,837 for Stomach. *indicates the significance level at which BiomedParse outperforms the best-competing method, with two-sided paired t-test p-value $< 1 \times 10^{-2}$ for **, p-value $< 1 \times 10^{-3}$ for ***, p-value $< 1 \times 10^{-4}$ for ****. Exact p-values for the comparison between BiomedParse and SegVol are as follows: p-value $< 2.23 \times 10^{-308}$ for All, p-value $< 1.86 \times 10^{-58}$ for Aorta, p-value $< 1.73 \times 10^{-7}$ for Bladder, p-value $< 3.44 \times 10^{-86}$ for Duodenum, p-value $< 5.00 \times 10^{-185}$ for Esophagus, p-value $< 3.37 \times 10^{-15}$ for Gallbladder, p-value $< 6.28 \times 10^{-99}$ for Inferior vena cava, p-value $< 5.08 \times 10^{-10}$ for Left adrenal gland, p-value $< 9.26 \times 10^{-31}$ for Left kidney, p-value $< 3.31 \times 10^{-37}$ for Liver, p-value $< 2.27 \times 10^{-56}$ for Pancreas, p-value $< 1.01 \times 10^{-16}$ for Right adrenal gland, p-value $< 2.98 \times 10^{-20}$ for Right kidney, p-value $< 1.09 \times 10^{-20}$ for Spleen, and p-value $< 4.68 \times 10^{-25}$ for Stomach.

https://doi.org/10.1038/s41592-024-02499-w



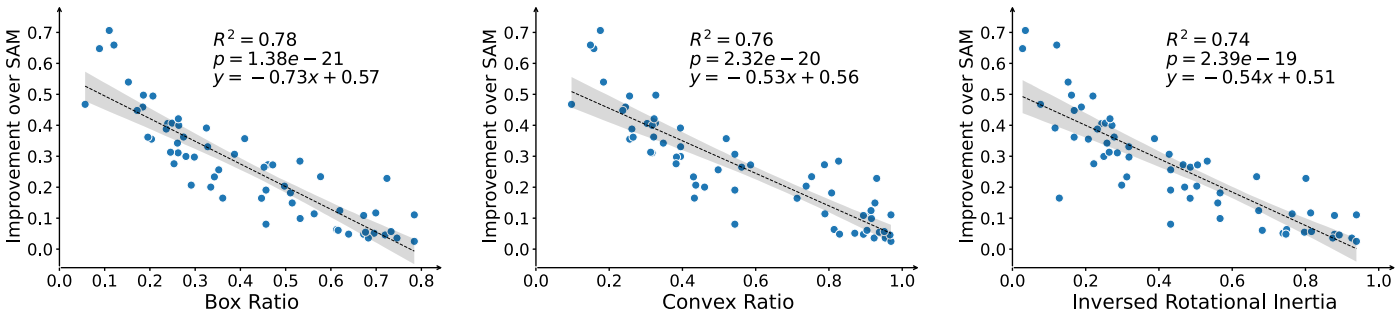
Extended Data Fig. 5 | Comparing BiomedParse with fine-tuned SAM and MedSAM. Bar plot comparing BiomedParse and SAM and MedSAM when SAM and MedSAM are both further trained on the entire BiomedParseData. Both SAM and MedSAM were provided with oracle bounding box around the segmentation target during the training and the inference stage. Each bar shows the mean of the distribution, with error bar indicating the 95% confidence interval. Each modality category contains multiple object types. Each object type was aggregated as the instance median to be shown in the plot. We show the numbers of object types in each modality are as follows. The numbers of object types in each modality are as follows: n = 105 for All, n = 26 for CT, n = 34 for MRI, n = 6 for X-Ray, n = 24 for Pathology, n = 7 for Ultrasound, n = 2 for Fundus, n = 3 for Endoscope, n = 2 for Dermoscopy, and n = 1 for OCT. *indicates the significance level at which BiomedParse outperforms the best-competing method, with two-sided paired t-test p-value $< 1 \times 10^{-2}$ for **, p-value $< 1 \times 10^{-3}$ for ***, p-value $< 1 \times 10^{-4}$ for ****. Exact p-values for the comparison between BiomedParse and SAM-FT with oracle box prompt are as follows: p-value $< 1.78 \times 10^{-7}$ for All, p-value $< 2.02 \times 10^{-2}$ for CT, p-value $< 1.32 \times 10^{-2}$ for X-Ray, p-value $< 3.52 \times 10^{-8}$ for Pathology, and p-value $< 1.49 \times 10^{-2}$ for Ultrasound.

https://doi.org/10.1038/s41592-024-02499-w



Extended Data Fig. 6 | Comparison between BiomedParse and competing methods on the MedSAM benchmark. We evaluated MedSAM and SAM using the ground truth bounding box for the segmentation.For nnU-Net and DeepLabV3+, we reported the evaluation reported by MedSAM. Results are shown by imaging modality, with statistical significance comparison between BiomedParse and best-competing method MedSAM. Each box shows the quartiles of the distribution, with center as the median, minimum as the first quartile, and maximum as the third quartile. The whiskers extend to the farthest data point that lies within 2 times the inter-quartile range (IQR) from the nearest quartile. Data points that lie outside the whiskers are shown as fliers. Each modality category contains multiple object types. Each object type was aggregated as the instance median to be shown in the plot. The numbers of object types in each modality are as follows: n = 50 for All, n = 18 for CT, n = 15 for MRI, n = 6 for X-Ray, n = 1 for Pathology, n = 6 for Ultrasound, n = 2 for Fundus, n = 1 for Endoscope, and n = 1 for Dermoscopy. * indicates the significance level at which BiomedParse outperforms the best-competing method, with two-sided paired t-test p-value $< 1 \times 10^{-2}$ for **, p-value $< 1 \times 10^{-3}$ for ***, p-value $< 1 \times 10^{-4}$ for ****. Exact p-values for the comparison between BiomedParse and MedSAM with oracle box prompt are as follows: p-value $< 2.98 \times 10^{-3}$ for All, p-value $< 7.08 \times 10^{-3}$ for CT, and p-value $< 4.35 \times 10^{-2}$ for MRI.

https://doi.org/10.1038/s41592-024-02499-w



Extended Data Fig. 7 | Comparing the improvement of BiomedParse over SAM with shape irregularity. Scatter plots comparing the improvement of BiomedParse over SAM with shape irregularity in terms of box ratio (left), convex ratio (middle), and inversed rotational inertia (right). Each dot represents the mean statistics over one object type in our segmentation ontology. We show the regression plot with the 95 confidence interval as the error bands. The p-values show the two-sided Wald test results.

nature portfolio

Corresponding author(s): Hoifung Poon
Sheng Wang
Mu Wei

Last updated by author(s): Sep 15, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All source image data were from publicly available datasets. We used Python (version 3.10.12) to curate and preprocess the image data. For the textual description for the objects in the images, we used GPT-4 provided by Azure OpenAI to generate text data.
Data analysis	This work uses open source code bases and libraries to analyze the data. We used SEEM(https://github.com/UX-Decoder/Segment-Everything-Everywhere-All-At-Once) for the main model architecture and training of the model on the datasets. Our model can be accessed at https://aka.ms/biomedparse-release, including the model weights and relevant source code. We include detailed methods and implementation steps in the Methods to allow for independent replication. We used matplotlib==3.8.2 and seaborn==0.11.2 to visualize the data. All the codes to reproduce our experiments will be made public upon publication.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

BiomedParseData can be accessed at https://aka.ms/biomedparse-release. The three real-world pathology images, including the annotations by pathologists and BiomedParse, can be accessed at https://aka.ms/biomedparse-release. BiomedParse can be accessed at https://aka.ms/biomedparse-release, including the model weights and rel-evant source code.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Our dataset was created by synthesizing 45 publicly available biomedical image segmentation datasets, comprising 1,050,227 images, 3,367,496 image-mask-label triples, and 6,696,943 image-mask-description triples. For model development, we used 340,081 images, 844,652 image-mask-label triples, and 1,729,052 image-mask-description triples. 102,855 image-mask-label triples were used to evaluate model performance. We decided on our data collection based on medical image segmentation literatures, and covered most of the common imaging modalities and tasks. We aimed for diversity of our coverage, and achieved a total data size that is among the largest so far.
Data exclusions	To ensure the quality of data, we imposed stringent inclusion criteria: each image had to be manually or semi-manually segmented at the pixel level, and a name was available for each segmented object from the dataset description. We excluded datasets that duplicate the tasks from other datasets in use to make sure the model training is balanced. We also excluded labels in the datasets that are noisy to ensure the quality of training.
Replication	We trained the model with randomization in gradient descent, batch data sampling and mixing, and data loader shuffling. The model performance is repeatable across different randomization. We evaluated the model with different randomization of text prompt, and achieved statistically in-differentiable performance. While the results are reproducible statistically, exact values will be affected by randomization and computational error in different environments.
Randomization	For model training and evaluation, we randomly split each original dataset into 80% training and 20% testing. Slices from each 3D volume always appear in the same split to prevent information leakage.
Blinding	Performance on the test image and labels were blind to the researchers before final evaluation. Any meta data of the source datasets were also held-out during model training.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging