Atlas is Your Perfect Context: One-Shot Customization for Generalizable Foundational Medical Image Segmentation

Ziyu Zhang[1⋆], Yi Yu[2⋆], Simeng Zhu[2], Ahmed H Aly[2], Yunhe Gao[3], Ning Gu[1(✉)], and Yuan Xue[2(✉)]

[1] Nanjing University, Nanjing, China
[2] The Ohio State University, Columbus, USA
Yuan.Xue@osumc.edu
[3] Stanford University, Stanford, USA
https://github.com/yuyi1005/AtlasSegFM

Abstract. Accurate anatomical structure segmentation in medical image is essential for diagnosis and treatment planning. While recent interactive segmentation foundation models enhance generalization through large-scale multimodal pretraining, they still depend on precise prompts and can fail in underrepresented clinical contexts (e.g., small organs-at-risk). We present AtlasSegFM, an atlas-guided framework that customizes off-the-shelf foundation models to new clinical contexts with a single annotated example. AtlasSegFM **1)** performs atlas-query registration to generate context-aware prompts, **2)** refines the segmentation with a frozen foundation model, and **3)** applies a lightweight adaptive fusion module to combine atlas priors with foundation-model inputs and predictions. Extensive experiments on six public and in-house datasets across radiotherapy and vascular scenarios show consistent gains, with the largest improvements on small and delicate structures. AtlasSegFM provides a lightweight, deployable solution for one-shot customization of segmentation foundation models in real-world clinical workflows.

1 Introduction

Interactive segmentation has emerged as a promising approach in medical imaging, enabling precise and customizable delineation by incorporating user input [14,21,22,29,47]. Unlike end-to-end methods, which may struggle to generalize across diverse anatomical structures and imaging modalities, interactive segmentation allows clinicians to guide the model so that results align with clinical expectations [13,39,46]. In clinical practice, segmentation demands vary widely such as abdominal organ delineation, vascular segmentation in angiography, and organs-at-risk identification in radiotherapy, each with distinct anatomical complexity, task-specific requirements, and imaging protocols. Moreover, physicians may differ in how they define and prioritize structures, particularly for small yet

⋆ Z. Zhang and Y. Yu—Equal contribution.

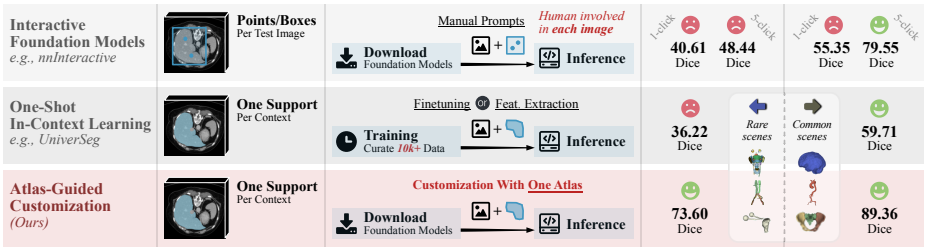

Fig. 1: Comparison of three segmentation paradigms. Interactive foundation models require per-image human prompts (points/boxes). One-shot in-context learning typically requires training on curated labeled datasets to learn cross-task prompting. Our AtlasSegFM adapts off-the-shelf foundation models to a target context using one annotated atlas. "Rare-scene" denotes contexts that are rare or not covered by the foundation model's reported training distribution in our evaluation (e.g., organs-at-risk and lower-limb vessels; see Sec. 4).

clinically critical regions, making it essential for segmentation tools to accommodate individual judgment. The ability to perform segmentation on demand, customized to specific clinical contexts or physician instructions, is therefore crucial for rare conditions, patient-specific variations, or limited annotated data, underscoring the need for flexible, adaptive, and clinically intuitive solutions in medical image analysis.

Existing customized segmentation typically relies on two main approaches: interactive segmentation and In-Context Learning (ICL). Interactive segmentation leverages user inputs, such as scribbles or clicks, to iteratively refine predictions and adapt to specific cases. While methods like MedSAM2 [22], nnInteractive [13], and ScribblePrompt [40] demonstrate its potential, they often require multiple interactions and rely heavily on the quality of user inputs. In contrast, ICL enables segmentation by providing contextual examples or prompts during inference, without the need for interaction or user input. ICL methods such as UniverSeg [2] and Iris [5] have shown the ability to generalize across various tasks and datasets by leveraging in-context examples, using reference examples or prompts to segment for specific scenarios.

As shown in Fig. 1, interactive methods offer flexibility by leveraging user inputs (e.g., points or boxes) to iteratively refine predictions during inference, whereas one-shot in-context learning methods use a single support example per context and require no interaction at inference. However, current ICL methods require large, carefully annotated datasets and training or extensive fine-tuning, making them both resource-intensive and time-consuming [2, 5, 28, 37]. Meanwhile, the growing availability of segmentation foundation models suggests an alternative strategy: instead of training a new model for each task, we can customize strong pretrained models to each clinical context.

Moreover, interactive foundation models struggle when prompts are ambiguous or when the target structure is rare or small [13, 21]. As shown in Fig. 2, interactive methods perform well in common scenarios, such as whole brain seg-

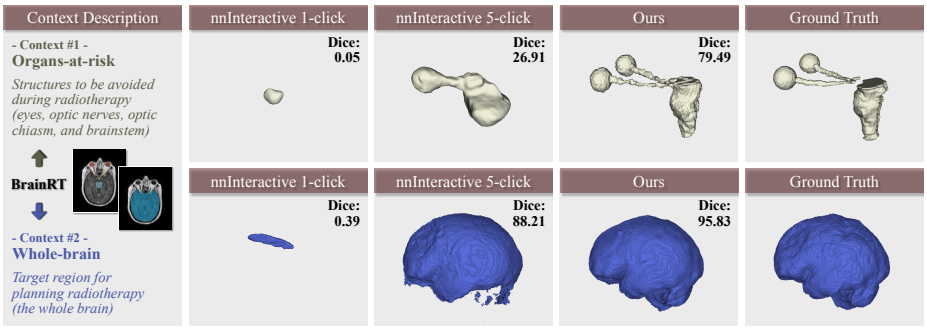

Fig. 2: Visual comparisons with recent interactive method (i.e., nnInteractive [13]) on the BrainRT dataset for radiotherapy. nnInteractive performs far below expectation in context #1 (organs-at-risk), a rare scenario that is underrepresented in its training data, whereas our method remains robust in both contexts.

mentation, achieving a Dice of 88.21% with 5-click interaction. However, their effectiveness drops significantly in delicate structures, with the Dice falling to 26.91% for organs-at-risk, where the targets are small, delicate structures embedded in complex context. This gap reflects both the limitations of pre-training data coverage and the absence of explicit contextual information in spatial prompts, which impedes generalization to unseen categories and fine-grained anatomy.

To address the above challenges, we propose a novel pipeline that combines structural priors of classical atlas-based segmentation with the representational power of modern segmentation foundation models. Our method leverages atlas-based registration to avoid the training burden typically associated with ICL while preserving critical structural priors in a single support example. The atlas mask serves as a context-aware prompt for interactive or non-interactive foundation models, providing more meaningful guidance and improving segmentation performance. This design enables efficient, contextually rich, and robust segmentation across diverse clinical scenarios. The main contributions are as follows:

- We investigate one-shot customization of segmentation foundation models for generalizable anatomical structure segmentation, where a frozen foundation model is adapted to a new clinical context using a single annotated atlas without requiring a dedicated training set.
- We present AtlasSegFM, a simple yet effective pipeline that combines the global anatomical consistency of atlas registration with the local refinement capability of segmentation foundation models. A lightweight adaptive fusion module integrates their complementary strengths at test time.
- Evaluated on multiple public and in-house datasets spanning organs-at-risk, pelvic anatomy, and vascular trees, AtlasSegFM consistently improves segmentation accuracy and boundary quality, particularly in underrepresented anatomical contexts where existing methods often fail, demonstrating strong out-of-distribution generalization.

2 Related Work

2.1 Medical Image Segmentation

Medical image segmentation has long been an active research area, with deep learning emerging as the dominant approach. Early deep learning methods relied on supervised learning, leading to the development of numerous architectures, including U-Net variants [30,50], Transformer-based models [6,35,49], and frameworks based on Mamba [16,31,43]. These methods have demonstrated remarkable performance across a wide range of segmentation tasks, particularly when sufficient annotated data is available. By incorporating an automated pipeline, nnU-Net [12] optimizes preprocessing, architecture, and hyperparameters, providing a robust and adaptable solution. Its comprehensive design has set a standard for supervised learning methods in the field.

However, these methods face significant challenges when dealing with out-of-distribution data or new segmentation tasks [45]. Their performance deteriorates in scenarios that deviate from the training domain, and adapting to new or specific tasks often requires retraining the model from scratch. This process is heavily reliant on annotated data and substantial computational resources, making it impractical for deployment in highly customized clinical tasks.

2.2 Segmentation Foundation Models

Foundation models have shown remarkable success in medical image segmentation [14,29]. Some are task-specific, such as vesselFM [38], a general-purpose zero-shot 3D vascular segmentation model with strong generalization capability. To enable user-customized segmentation, another line of research has explored interactive approaches, such as MedSAM [21,22] and nnInteractive [13]. By incorporating user prompts (e.g., clicks and bounding boxes), these methods simplify the annotation and achieve strong performance with minimal user interaction [4,7,8,34,40].

While these models leverage powerful pretraining on large datasets to generalize across tasks, they still face two challenges: **1)** Prompts are often ambiguous, and the models infer user intent based on prior training data. Consequently, in rare contexts (e.g., organs-at-risk in Fig. 2), their performance falls short of expectations. **2)** They require repeated user interactions for each image, making them inefficient for large-scale segmentation tasks or routine clinical workflows, where many patients follow similar procedures. In-context learning is a possible solution to these challenges.

2.3 In-Context Learning Methods

In-context learning allows models to generalize to unseen tasks by leveraging a context set of labeled image-segmentation pairs at inference time [10,39,41,42,48]. Few-shot segmentation methods [2,20], primarily designed to reduce labeling effort, were explored first. However, most of these methods rely on small models

and limited data, and some [3, 15, 25, 32] are further fine-tuned on the target domain (referred to as cross-domain ICL). As a result, they lack the generality needed to perform arbitrary segmentation like foundation models. Therefore, some ICL studies train on larger datasets, including UniverSeg [2] and Iris [5].

With the growing availability of pretrained foundation models (see Sec. 2.2), a key limitation of these ICL approaches has become evident—they must be trained from scratch or fine-tuned on specific datasets, preventing them from leveraging the continuously improving foundation models. To address this, some studies have incorporated foundation models into ICL [44]. Their application to the medical domain is still at an early stage [17, 23, 51], with either unavailable code or limited performance. Moreover, existing ICL methods are primarily designed and evaluated on 2D images, limiting their applicability to 3D segmentation. How to fully leverage the capabilities of foundation models in ICL remains largely unexplored in the literature.

2.4 Atlas-Based Segmentation

Atlas-based methods represent a classical approach that does not require annotations or learning [11]. These methods treat segmentation as a registration problem, where the atlas is aligned to unlabeled data using optimization algorithms. Through this alignment, the structural information from the atlas can be transferred to the target image, enabling segmentation without direct supervision. Historically, atlas-based approaches iteratively minimize a cost function measuring the similarity between atlas and query [36]. More recently, some deep-learning-based registration methods have been proposed [1, 9, 33]. By leveraging neural networks, these approaches efficiently learn the mappings from atlas to query, significantly reducing computational overhead while improving accuracy.

3 Methods

Medical image segmentation aims to assign a label to each pixel (or voxel) of an input image $X \in \mathbb{R}^{H\times W\times D}$ (e.g., CT/MRI scans) based on its underlying anatomical structure. Given a segmentation model $f(\cdot)$, the goal is to predict a label map $Y \in \mathbb{R}^{H\times W\times D}$, where each voxel $y_i \in Y$ corresponds to a predefined anatomical or pathological category, as $Y = f(X)$.

In clinical practice, segmentation often needs to be *Customized* for specific tasks, such as segmenting rare structures or adapting to new imaging modalities. Inspired by ICL, where a support set is used to provide domain-specific guidance, a contextualized segmentation task can be expressed as:

$$Y_{query} = f(X_{query}, \mathcal{C}) \tag{1}$$

where the context support $\mathcal{C}$ provides task- and anatomy-specific guidance. In AtlasSegFM, $\mathcal{C} = (X_{support}, Y_{support})$ is a a single annotated atlas (one support pair) for the target structure (or a target class in multi-class datasets) and X_{query} is the query image.

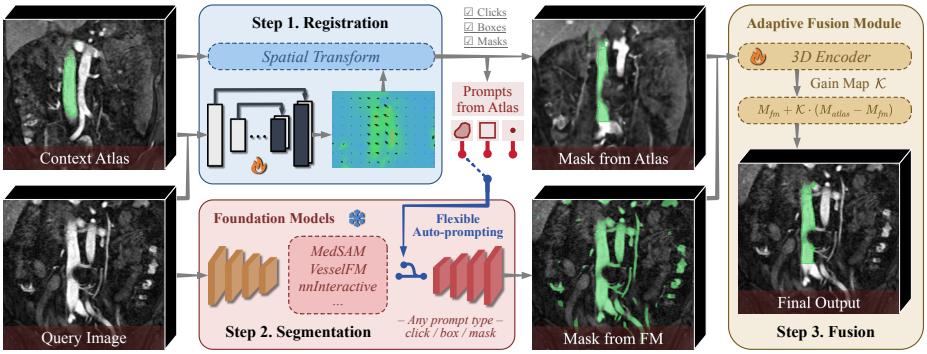

Fig. 3: Pipeline including three steps: **1)** Registration between query and support to obtain the mask (see Sec. 3.1), **2)** Prompting the foundation model based on the mask from atlas (see Sec. 3.2), and **3)** Fusion of the two masks to obtain the final result (see Sec. 3.3). Our model uses an inference-only design, where the "fire" denotes test-time adaptation and the "snow" remains frozen.

Existing ICL methods rely on feature similarity between $X_{support}$ and X_{query}, which either requires extensive training and large-scale labeled data (e.g. UniverSeg [2], Iris [5]) or depend on large pretrained encoders such as DINO [24] for feature extraction. As shown in Fig. 3, we propose an atlas-guided, in-context segmentation framework that combines the strengths of classical atlas-based registration and advanced foundation models. AtlasSegFM consists of three steps: **1)** Atlas-based registration to provide context-aware prompts, **2)** Segmentation refinement using foundation models with these prompts, and **3)** Fusion module to integrate predictions from both sources.

3.1 Registration and Prompt Generation

To generate task-specific prompts for foundation models, we adopt an atlas-based registration approach as the initialization step for in-context learning with two main advantages: **1)** atlas-based registration does not require training or labeled datasets, making it lightweight and easy to deploy in clinical workflows, and **2)** it inherently leverages structural priors, which are particularly valuable for medical image segmentation tasks where anatomical consistency is critical.

Traditional atlas-based segmentation relies on iterative optimization to align the atlas to the query image, which is computationally expensive and time-consuming. To overcome this, we develop a test-time registration network derived from RDP [33], utilizing its network architecture and optimization objectives with adaptations for test-time optimization. Formally, given a context atlas (X_{atlas}, Y_{atlas}) and a query image X_{query}, we estimate the spatial transformation T by optimizing:

$$T^{\star} = \arg\min_{T} \mathcal{L}_{reg}(T(X_{atlas}), X_{query}) \tag{2}$$

where $\mathcal{L}_{reg}$ is a similarity-based registration loss, such as normalized cross-correlation or mutual information. The registered atlas $T^{\star}(X_{atlas})$ and its corresponding label $T^{\star}(Y_{atlas})$ yield a coarse segmentation mask M_{atlas}, which serves as a structural prior.

Original RDP optimizes registration by training on large datasets to learn network parameters. In AtlasSegFM, we repurpose it as a test-time optimization framework, where the network parameters are optimized on-the-fly for each specific atlas-query pair during inference. In addition, we perform rigid and affine pre-registration prior to network optimization, ensuring coarse alignment and enhancing final registration accuracy (as validated in the ablation Table 4).

After generating the coarse segmentation mask M_{atlas} from the registered atlas, we adapt it as the prompt context for foundation models. Specifically, our framework can provide three different types of prompts:

- Click: The centroid of the largest connected region within M_{atlas} is extracted as the click prompt P_{click}.
- Box: The minimum circumscribed bounding box that encloses M_{atlas} at the middle slice along z-axis is computed as the box prompt P_{box}.
- Mask: The segmentation mask M_{atlas} is directly used as the mask prompt P_{mask} for foundation models that support mask prompting.

These atlas-derived prompts inject explicit anatomical context (location, extent, and plausible shape) into the foundation model. Additional quantitative evidence comparing prompt types is reported in Sec. 4.2.

Following the inference protocols of nnInteractive [13] and MedSAM2 [22], point and box prompts are provided on one slice. Specifically, we derive the 2D bounding box from the middle slice of M_{atlas}. nnInteractive utilizes this input to directly infer the 3D volume, whereas MedSAM2 initiates segmentation at the middle slice and propagates the prediction bidirectionally through the volume.

3.2 Foundation Model Inference with Auto-Prompts

To seamlessly integrate different foundation models, our pipeline is designed to support both interactive and non-interactive foundation models. For promptable models, prompts generated in the previous step, denoted as P_i with $i \in \{click, box, mask\}$, are used to guide the segmentation refinement. The foundation model can select any supported prompt type (i.e., click, box, or mask) according to its interface, enabling flexible integration across models with different prompting mechanisms. For foundation models that do not support prompting (i.e. vesselFM [38]), the same pipeline simply feeds the input images directly to the foundation model without prompts. In our experiments with nnInteractive and MedSAM2, we use atlas-derived mask prompts by default, and analyze click/box/mask prompting in Sec. 4.2.

This unified auto-prompting design allows our pipeline to flexibly adapt to different foundation models and scenarios, leveraging prompting when available

while remaining compatible with models that operate without prompts. The foundation model output can be expressed as:

$$M_{fm} = f_{FM}(X_{query}, P_i) \tag{3}$$

where f_{FM}represents the foundation model and P_i is the prompt from atlas.

3.3 Atlas-FM Adaptive Fusion

Although the foundation model provides fine-grained segmentation M_{fm}, it may occasionally fail in ambiguous regions or under challenging scenarios. On the other hand, the atlas-based mask M_{atlas} offers robust structural priors but generally lacks fine detail. To address the limitations of both methods, we design a fusion module to dynamically integrate the predictions from the atlas and the foundation model, leveraging their complementary strengths.

The final mask M_{final} is computed by combining M_{atlas} and M_{fm} through a voxel-wise reliability-gated fusion:

$$M_{final} = M_{fm} + \mathcal{K} \cdot (M_{atlas} - M_{fm}) \tag{4}$$

where $\mathcal{K}$ is a spatial gain map that selects which source to trust at each voxel. A larger $\mathcal{K}$ indicates higher local reliability of the atlas prediction relative to the FM prediction, whereas a smaller $\mathcal{K}$ makes the output rely more on the FM.

Rather than using a fixed or global fusion weight, we estimate $\mathcal{K}$ using a lightweight adaptive reliability estimator. The estimator takes the query image X_{query} and the two predictions M_{atlas} and M_{fm} as inputs, and predicts a voxel-wise reliability gate as:

$$\mathcal{K} = \sigma(g([X_{query}, M_{atlas}, M_{fm}])) \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid activation function to ensure $\mathcal{K} \in [0, 1]$. In practice, $g(\cdot)$ first derives a feature volume from the normalized query image, the logit representations of M_{fm} and M_{atlas}, their signed logit difference, the voxel-wise disagreement between the two soft masks, and the entropy maps of both predictions. These cues encode local anatomical context, mask conflict, and uncertainty, enabling the network to estimate which source is more reliable at each voxel during test-time adaptation. The feature volume is then processed by a lightweight 3D encoder followed by a sigmoid activated gate. The resulting gate adaptively controls the fusion between the foundation-model prediction and the atlas prediction. The estimator parameters are optimized at test time using a cycle-transformation loss, as detailed in the implementation details.

4 Experiments

Datasets. We evaluate our proposed method on six datasets, covering commonly used few-shot medical image segmentation tasks and clinically significant challenges. To ensure a fair evaluation of generalization, all six datasets (public and in-house) are excluded from the nnInteractive training corpus, and we use fully held-out splits for all experiments.

- **HaN-Seg** [26]: A head and neck dataset contains 42 CT scans annotated for organs-at-risk. It serves as a standard benchmark for radiotherapy planning, presenting significant challenges in segmenting densely packed and low-contrast anatomical structures within the complex head and neck region.
- **SegRap** [19]: Targeting segmentation for radiotherapy in nasopharyngeal carcinoma, this dataset contains 30 CT scans annotated for organs-at-risk, many of which are distorted or affected by nearby tumors.
- **Pengwin** [18]: A pelvic bone segmentation dataset containing 100 CT scans annotated for various pelvic anatomical structures. It presents unique challenges in delineating complex bone geometries, articulating joints, and varying bone densities across different patients.
- **AVT** [27]: Aortic vessel tree dataset contains 56 CTA scans for vascular segmentation with segmentation masks of the aorta and its branches.
- **Fe-MRA**: An in-house dataset of 50 ferumoxytol-enhanced MRA scans of the lower extremities, with arterial and venous structures annotated into 12 fine-grained categories such as great saphenous veins and femoral arteries. It targets lower-limb vasculature in clinical scenarios including diagnosis and treatment planning for varicose veins and arterial thrombosis.
- **BrainRT**: An in-house dataset of 60 CT/MR scans for brain tumor radiotherapy planning, where radiologists delineated organs on CT and labels were registered to MR. It covers two crucial segmentation contexts: whole brain and organs-at-risk (eyes, optic nerves, optic chiasm, and brainstem), whose delineation is essential to minimize radiation-induced damage.

Protocol and metrics. Following [25], we adopt a 5-fold cross-validation protocol, where one sample from the remaining validation set is used as the context support for testing the query images in each fold. For datasets with multiple categories, each category is treated as a segmentation context. All data preprocessing procedures strictly follow the standard pipeline established by MedSAM [21]. Four metrics are used: **1)** Dice measuring the overlap between the prediction and ground truth; **2)** NSD (Normalized Surface Dice) measuring the agreement between two surfaces within a specified tolerance; **3)** HD (Hausdorff Distance 95) quantifying the boundary error by computing the 95th percentile of the Hausdorff distance, lower better; **4)** clDice (centerline Dice) emphasizing the correctness of the centerline connectivity, especially suitable for vessel.

Baselines. We compare our method against three categories: **1)** Supervised learning methods [12] directly trained on the target dataset; **2)** Foundation models [13, 22], where prompts are sampled from the ground truth mask during inference to guide the segmentation; **3)** In-context learning methods [2, 5, 28, 37] that do not require training or fine-tuning on the target dataset, relying on general-purpose pretrained models or inference-time adaptation. For promptable foundation-model baselines, we follow the standard evaluation protocol and sample point prompts from the query ground-truth masks. This provides an optimistic oracle localization signal that is unavailable in deployment. In contrast, AtlasSegFM derives its prompts automatically from only the support atlas and the unlabeled query image through atlas-query registration.

Table 1: Comparisons of organs-at-risk segmentation performance on the HaN-Seg and SegRap datasets, and pelvic bone segmentation performance on the Pengwin dataset.

Methods	HaN-Seg		SegRap		Pengwin	
	Dice↑	NSD↑	Dice↑	NSD↑	Dice↑	NSD↑
nnUNet (supervised upper bound) [12]	75.10	82.48	80.20	78.18	98.90	99.10
▼ *Foundation Models (1 or 5 clicks per test image)*						
MedSAM2-5 (2025) [22]	28.68	35.45	26.34	41.88	92.67	94.76
nnInteractive-1 (2025) [13]	42.42	52.08	45.67	56.84	72.09	75.29
nnInteractive-5 (2025) [13]	**55.29**	**66.24**	**50.02**	**69.32**	**93.63**	**96.00**
▼ *In-context Models (One support each context)*						
Register-based[1] (TMI 2024) [33]	48.72	67.96	44.50	66.95	82.75	85.25
SegGPT (ICCV 2023) [37]	24.52	31.26	27.31	42.65	34.20	37.52
UniverSeg (ICCV 2023) [2]	35.49	45.33	39.28	56.86	56.10	63.03
Tyche (CVPR 2024) [28]	33.99	43.10	37.45	54.22	42.74	46.90
Iris (CVPR 2025) [5]	37.57	48.28	42.73	60.07	68.14	71.75
Ours (MedSAM2-based)	52.61	65.74	49.45	69.19	94.86	96.74
Ours (nnInteractive-based)	**63.71**	**77.80**	**69.19**	**87.39**	**95.50**	**96.96**

[1] Using RDP [33] to estimate the deformation field from the support to the query and apply it to the mask of support to obtain the mask of query.

Implementation details. Our method consists of two learnable components, which are test-time optimized: **1)** Atlas-registration module. We employ the Recursive Deformable Pyramid (RDP) network [33], which utilizes a dual-stream encoder and a pyramid decoder to recursively predict deformation fields from coarse to fine. It is optimized between support and query images using the Normalized Cross Correlation (NCC) loss with a learning rate of 1e-4 for 300 iterations. **2)** Adaptive fusion module. The fusion estimator $g(\cdot)$ is optimized at test time using a cycle-transformation Dice loss. Specifically, after support-to-query registration, the warped support mask is used as the atlas-derived prompt for the frozen foundation model, and the fused prediction is produced in the query space. We then warp the fused prediction back to the support space and compute the Dice loss against the available support label. No query ground-truth mask is used during this optimization, and this is not a direct comparison between the support label and itself because the prediction has undergone deformation, foundation-model refinement, adaptive fusion, and inverse warping. $g(\cdot)$ is updated for 100 iterations with a learning rate of $1e^{-5}$, while the foundation model remains frozen. For vascular segmentation, a recent foundation model vesselFM [38] is adopted, whereas nnInteractive [13] is applied to other scenarios using mask-based prompts.

4.1 Comparisons with State-of-the-art

Results on organs-at-risk and pelvic bone datasets. Table 1 summarizes the Dice and NSD score comparisons across different methods on the HaN-Seg [26], SegRap [19], and Pengwin [18] datasets, focusing on the segmentation of organs-at-risk and pelvic bones. Our method (nnInteractive-based) achieves the highest mean Dice scores among all in-context models across all datasets, with 63.71% on HaN-Seg, 69.19% on SegRap, and 95.50% on Pengwin, outper-

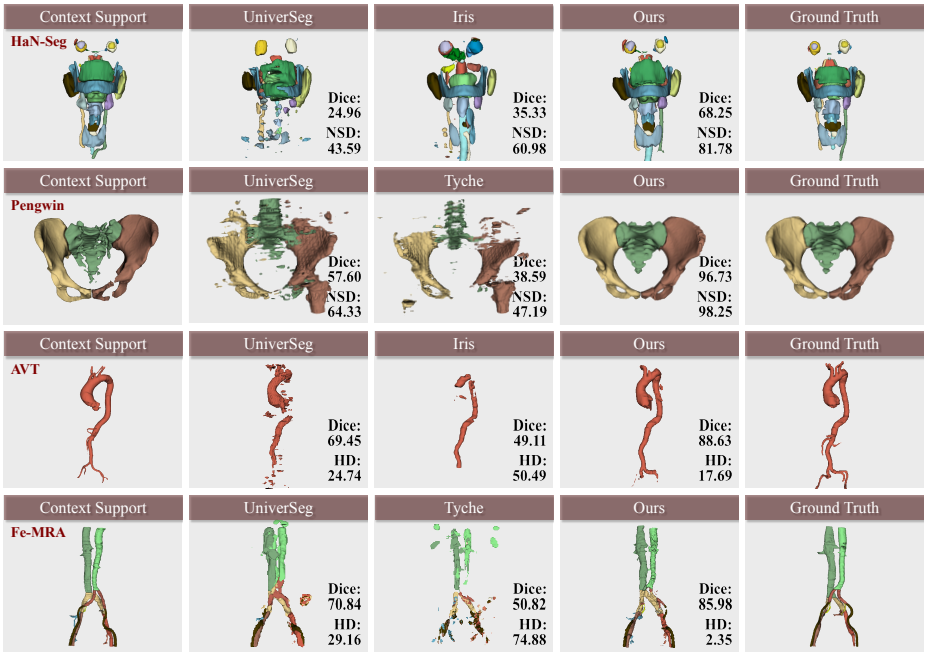

Fig. 4: Visual comparisons against recent one-shot learning methods with open-source implementations (i.e., UniverSeg [2], Tyche [28], and Iris [5]) across four datasets. Each color corresponds to a context, and all contexts are combined in the visualization.

forming all baseline methods. Our approach achieves substantial performance gains over recent ICL methods, improving the mean Dice score on HaN-Seg by 26.14% compared to Iris [5] and 29.72% compared to Tyche [28], highlighting its robust generalization and adaptability across diverse scenarios. Even compared to supervised learning methods, our approach achieves competitive performance, delivering segmentation quality comparable to nnUNet [12] (e.g., 95.50% vs. 98.90% on Pengwin), which relies on full supervision and target-specific training. This highlights the effectiveness of our method in achieving high-quality segmentation without the need for extensive labeled data.

Qualitative results in Fig. 4 (first two rows) further validate our method's capability to produce segmentations with clear boundaries and fine structural details. In contrast, others often struggle with incomplete or inconsistent outputs due to limited structural priors or contextual information.

Results on brain datasets. Table 2 reports segmentation performance on the BrainRT dataset, where our method achieves the highest Dice and NSD scores across all metrics for both tasks. For whole-brain segmentation, while most methods achieve reasonable performance, our approach stands out by delivering the highest accuracy with consistently smooth and anatomically precise segmentation, achieving a Dice score of 91.24% and an HD of 6.78. For the highly challenging organs-at-risk segmentation (e.g., optic nerves and brainstem), our method

Table 2: Comparisons of the brain-structure and organs-at-risk segmentation performance on the BrainRT dataset.

Methods	*Whole-brain* ← **BrainRT** → *Organs-at-risk*					
	Dice↑	NSD↑	HD↓	Dice↑	NSD↑	HD↓
▼ *Foundation Models (1 or 5 clicks per test image)*						
nnInteractive-1 (2025) [13]	0.62	1.82	104.04	30.65	24.38	90.01
nnInteractive-5 (2025) [13]	**61.63**	**40.31**	**27.62**	**39.09**	**26.42**	**59.63**
▼ *In-context Models (One support each context)*						
SegGPT (ICCV 2023) [37]	50.60	58.13	19.38	2.61	3.60	84.73
UniverSeg (ICCV 2023) [2]	83.32	32.44	15.40	0.39	3.83	32.59
Tyche (CVPR 2024) [28]	88.96	47.10	14.69	4.37	10.03	38.76
Iris (CVPR 2025) [5]	90.28	59.47	8.56	10.45	19.95	44.52
Ours (nnInteractive-based)	**91.24**	**63.08**	**6.78**	**77.07**	**77.55**	**5.17**

Table 3: Comparisons of the vessel segmentation on the AVT and Fe-MRA datasets.

Methods	AVT			Fe-MRA		
	Dice↑	clDice↑	HD↓	Dice↑	clDice↑	HD↓
▼ *Foundation Models (1 or 5 clicks per test image)*						
vesselFM (CVPR 2025) [38]	28.19	7.11	274.02	**60.31**	**41.74**	124.34
nnInteractive-1 (2025) [13]	63.31	53.27	65.79	43.69	23.12	182.87
nnInteractive-5 (2025) [13]	**83.38**	**70.52**	**43.69**	49.36	34.00	**110.01**
▼ *In-context Models (One support each context)*						
SegGPT (ICCV 2023) [37]	68.73	48.78	45.05	0.31	0.69	133.15
UniverSeg (ICCV 2023) [2]	39.71	31.89	51.12	69.71	53.20	39.61
Tyche (CVPR 2024) [28]	53.24	42.76	70.44	50.21	59.64	79.28
Iris (CVPR 2025) [5]	45.44	41.93	58.53	28.35	13.66	38.88
Ours (vesselFM-based)	**81.34**	**72.04**	**30.84**	**84.42**	**82.99**	**3.00**

achieves a Dice score of 77.07% and an NSD of 77.55%, significantly outperforming all baselines by a massive margin. Accurate segmentation of these structures is crucial in radiotherapy to minimize severe radiation-induced damage; however, even the state-of-the-art Iris method struggles with unseen structures, underscoring the importance and difficulty of this task. As shown in Fig. 2, foundation models (e.g., nnInteractive 1/5 clicks) perform poorly or even fail on organs-at-risk segmentation due to the intricate structures and small size of these regions, which demand finer detail preservation and better contextual understanding.

Results on vessel datasets. In Table 3, we evaluate on vessel segmentation, a challenging task due to intricate geometries and thin structures. Our method achieves the highest Dice scores (81.34% on AVT and 84.42% on Fe-MRA) and clDice scores (72.07% on AVT and 82.99% on Fe-MRA), while significantly outperforming other methods in HD (30.84 on AVT and 3.00 on Fe-MRA). These results demonstrate its capability to accurately capture complex vascular structures and maintain smooth, anatomically consistent surfaces.

As shown in Fig. 4 (last 2 rows), our method reconstructs the full topology of the aortic vessel tree on AVT and clearly delineates both arterial and venous structures on Fe-MRA, outperforming other methods that struggle with incomplete outputs. This highlights our method's ability to produce detailed, anatomically consistent segmentations.

Table 4: Ablation with incremental addition of modules on the Pengwin and Fe-MRA datasets.

Module Settings	Pengwin	HaN-Seg	Fe-MRA
Baseline #1 (Atlas)	63.24	37.32	35.73
+ Rigid pre-registration	70.55	41.89	66.43
+ Affine pre-registration	82.75	48.72	81.64
Baseline #2 (FM)	72.09*	42.42*	55.28
+ Prompts from atlas	94.56	61.74	N/A†
+ Fusion (Final)	**95.50**	**63.71**	**84.42**

*One point prompt is provided for nnInteractive.
†The foundation model vesselFM is not promptable.

Table 5: Inference time (min per image) and learnable parameters.

Methods	Inference Time	Learnable Param.
SegGPT	18.1	1.4G
UniverSeg	2.2	1.2M
Tyche	2.1	1.2M
Ours (Total)	1.8	8.6M
- Regist.	1.5	8.5M
- FM	0.01	-
- Fusion	0.3	0.1M

4.2 Ablation Studies

Table 4 shows the ablation studies, where we evaluate the impact of different modules in our framework on the Pengwin, HaN-Seg, and Fe-MRA datasets. This analysis highlights the contributions of atlas-based registration, pre-registration strategies, foundation model prompts, and the fusion module. These datasets are selected to cover three representative clinical scenarios with different anatomical properties and potential failure modes: relatively rigid bony structures in Pengwin, deformable head-and-neck organs at risk in HaN-Seg, and thin, topology-sensitive vascular structures in Fe-MRA.

Atlas-FM registration. The performance of Baseline #1, which uses only atlas-based registration, sets the starting point for analysis. By simply aligning between atlas and query, the Dice scores are 63.24% on Pengwin and 35.73% on Fe-MRA. Pre-registration also improves the performance, with the rigid one raising Dice scores to 70.55% on Pengwin and 66.43% on Fe-MRA. Further incorporating affine pre-registration to better account for anatomical variations boosts performance to 82.75% and 81.64% on Pengwin and Fe-MRA.

Foundation models. Baseline #2 evaluates the independent performance of the foundation model. For Pengwin, when we prompt nnInteractive with one point per image, the Dice score is 72.09%. On Fe-MRA, vesselFM achieves a Dice score of 55.28%, leveraging its pre-trained capabilities for fine-grained segmentation. By providing atlas-derived prompts to the foundation model, the performance on Pengwin increases to 94.56%, showcasing the utility of structural priors in guiding the foundation model.

Atlas-FM fusion. Atlas registration provides accurate global structure, whereas foundation models demonstrate better local details. As listed in the last row, our adaptive fusion module combines the registered atlas prediction and the foundation model prediction, achieving Dice scores of 95.50% on Pengwin and 84.42% on Fe-MRA, surpassing either prediction alone and demonstrating the benefit of fusing their complementary strengths.

Runtime and parameter efficiency. Table 5 measures the computational cost on NVIDIA 4090 GPU using a $(256 \times 256 \times 240)$ image from the BrainRT organs-at-risk task. Unlike slice-by-slice methods such as SegGPT, Tyche, and UniverSeg, our method operates directly on 3D images, enabling faster inference.

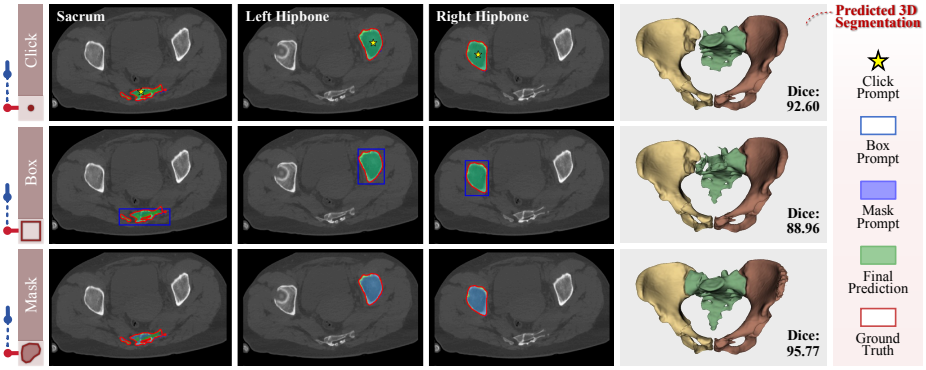

Fig. 5: Click, box, and mask prompts provided by our atlas guidance. Our method supports different foundation models/prompt types and achieves high accuracy across them, with mask prompts performing best (visualized with nnInteractive).

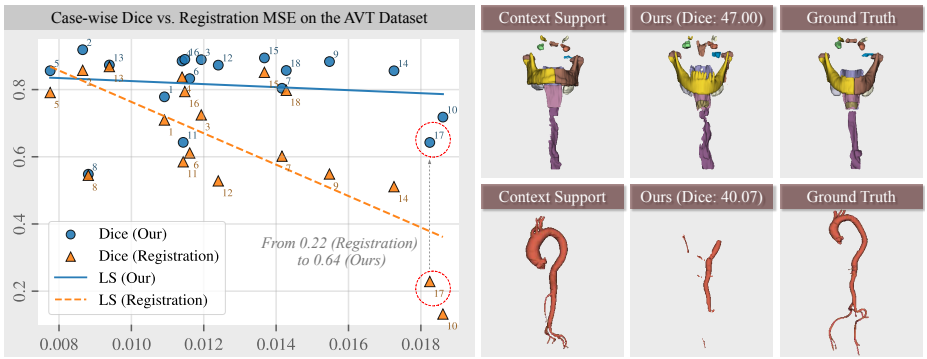

Fig. 6: Left: Comparison of the performance degradation under imperfect alignment. Right: Failure cases of our method. Top-right (from SegRap) erroneously merges the left and right mandibles. Bottom-right (from AVT) fails to fully recover the aortic arch.

It takes 1.8 minutes per image, compared to 18.1 minutes for SegGPT and 2.1 minutes for Tyche. The runtime mainly consists of atlas-based registration (1.5 minutes) and prediction fusion (0.3 minutes), whereas the foundation model inference takes 0.01 minutes. With only 8.6M learnable parameters and a single support-query pair for test-time adaptation, our method is lightweight, practical, and demonstrates relatively higher efficiency.

Different prompting types. We analyze click/box/mask prompts for nnInteractive in Fig. 5. On the Pengwin case, atlas-derived prompts yield 92.60 Dice (click), 88.96 Dice (box), and 95.77 Dice (mask), highlighting that dense mask prompts provide the most informative context when initialization is coarse.

Sensitivity to registration quality. We further analyze how segmentation performance changes with registration quality. As shown in Fig. 6 (left), Dice of registration degrades sharply as MSE increases, whereas our method shows

a flatter degradation trend. In the highlighted case, our method improves Dice from 0.22 to 0.64 after foundation-model refinement and adaptive fusion. This suggests that our method does not simply inherit the warped atlas and can partially compensate for imperfect alignment.

Failure cases. Despite its strengths, our method encounters challenges in certain difficult cases. As shown in Fig. 6 (right), it erroneously merges the left and right mandibles in the SegRap dataset and misses critical regions in the AVT. The former arises because the left and right mandibles are artificially separated during annotation without a clear anatomical boundary, while nnInteractive tends to segment the mandible as a single continuous structure. The latter occurs when the support and query images differ substantially, making it difficult for the matching network to regress accurate correspondences.

5 Conclusion

In this paper, we introduced AtlasSegFM, a simple yet effective medical image segmentation framework. By leveraging the globally consistent structure provided by the atlas and the representation power of foundation models, AtlasSegFM fuses their complementary strengths and achieves competitive accuracy. It offers the following advantages: **1)** It fully exploits off-the-shelf foundation models, enabling context-specific customization with a single atlas and no additional training data. **2)** Experiments across six datasets demonstrate an average Dice of 80.35%, substantially outperforming both uncustomized foundation models and existing in-context learning methods. **3)** In rare contexts underrepresented in foundation model pretraining, such as organs-at-risk, our method achieves a Dice of 73.60% without training on similar cases, demonstrating strong out-of-distribution robustness for complex, delicate structures.

Limitations and future direction. While effective for organ segmentation, our method is less suitable for lesion segmentation due to the dependence on structural consistency between support and query images, which focal lesions often lack. It also faces challenges with patients whose anatomy differs markedly from the atlas. Future work will explore multi-atlas strategies for greater robustness and refine prompt generation for more context-aware guidance.

Acknowledgements

Ziyu Zhang was supported in part by the Natural Science Foundation of Jiangsu Province of China (BK20241202), the China Postdoctoral Science Foundation (2024M751394, GZC20240691), and the Jiangsu Funding Program for Excellent Postdoctoral Talent (2024ZB433).

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging **38**(8), 1788–1800 (2019) 5

2. Butoi, V.I., Ortiz, J.J.G., Ma, T., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Universeg: Universal medical image segmentation. In: IEEE/CVF International Conference on Computer Vision. pp. 21438–21451 (2023) 2, 4, 5, 6, 9, 10, 11, 12
3. Cheng, Z., Wang, S., Xin, T., Zhou, T., Zhang, H., Shao, L.: Few-shot medical image segmentation via generating multiple representative descriptors. IEEE Transactions on Medical Imaging **43**(6), 2202–2214 (2024) 5
4. Du, Y., Bai, F., Huang, T., Zhao, B.: Segvol: Universal and interactive volumetric medical image segmentation. In: Conference on Neural Information Processing Systems. vol. 37, pp. 110746–110783 (2024) 4
5. Gao, Y., Liu, D., Li, Z., Li, Y., Chen, D., Zhou, M., Metaxas, D.N.: Show and segment: Universal medical image segmentation via in-context learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20830–20840 (2025) 2, 5, 6, 9, 10, 11, 12
6. Gao, Y., Zhou, M., Metaxas, D.N.: Utnet: a hybrid transformer architecture for medical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 61–71 (2021) 4
7. Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q.: 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. Medical Image Analysis **98**, 103324 (2024) 4
8. Guo, H., Zhang, J., Huang, J., Mok, T.C., Guo, D., Yan, K., Lu, L., Jin, D., Xu, M.: Towards a comprehensive, efficient and promptable anatomic structure segmentation model using 3d whole-body ct scans. In: AAAI Conference on Artificial Intelligence. pp. 3247–3256 (2025) 4
9. Hoffmann, M., Billot, B., Greve, D.N., Iglesias, J.E., Fischl, B., Dalca, A.V.: Synthmorph: learning contrast-invariant registration without acquired images. IEEE Transactions on Medical Imaging **41**(3), 543–558 (2021) 5
10. Hu, J., Shang, Y., Yang, Y., Guo, X., Peng, H., Ma, T.: Icl-sam: Synergizing in-context learning model and sam in medical image segmentation. In: Medical Imaging with Deep Learning. pp. 641–656 (2024) 4
11. Iglesias, J.E., Sabuncu, M.R.: Multi-atlas segmentation of biomedical images: a survey. Medical Image Analysis **24**(1), 205–219 (2015) 5
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021) 4, 9, 10, 11
13. Isensee, F., Rokuss, M., Krämer, L., Dinkelacker, S., Ravindran, A., Stritzke, F., Hamm, B., Wald, T., Langenberg, M., Ulrich, C., et al.: nninteractive: Redefining 3d promptable segmentation. arXiv preprint arXiv:2503.08373 (2025) 1, 2, 3, 4, 7, 9, 10, 12
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023) 1, 4
15. Lin, Y., Chen, Y., Cheng, K.T., Chen, H.: Few shot medical image segmentation with cross attention transformer. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 233–243 (2023) 5
16. Liu, J., Yang, H., Zhou, H.Y., Yu, L., Liang, Y., Yu, Y., Zhang, S., Zheng, H., Wang, S.: Swin-umamba†: Adapting mamba-based vision foundation models for medical image segmentation. IEEE Transactions on Medical Imaging (2024) 4
17. Liu, Y., Zhu, M., Chen, H., Wang, X., Feng, B., Wang, H., Li, S., Vemulapalli, R., Shen, C.: Segment anything in context with vision foundation models. International Journal of Computer Vision **133**(10), 7460–7485 (2025) 5

18. Liu, Y., Yibulayimu, S., Sang, Y., Zhu, G., Wang, Y., Zhao, C., Wu, X.: Pelvic fracture segmentation using a multi-scale distance-weighted neural network. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 312–321 (2023) 9, 10
19. Luo, X., Fu, J., Zhong, Y., Liu, S., Han, B., Astaraki, M., Bendazzoli, S., Toma-Dasu, I., Ye, Y., Chen, Z., et al.: Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. Medical Image Analysis **101**, 103447 (2025) 9, 10
20. Lv, J., Zeng, X., Wang, S., Duan, R., Wang, Z., Li, Q.: Robust one-shot segmentation of brain tissues via image-aligned style transformation. In: AAAI Conference on Artificial Intelligence. pp. 1861–1869 (2023) 4
21. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024) 1, 2, 4, 9
22. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025) 1, 2, 4, 7, 9, 10
23. Mao, X., Xing, X., Meng, F., Liu, J., Bai, F., Nie, Q., Meng, M.: One polyp identifies all: One-shot polyp segmentation with sam via cascaded priors and iterative prompt evolution. In: IEEE/CVF International Conference on Computer Vision. pp. 24182–24191 (2025) 5
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 6
25. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervised learning for few-shot medical image segmentation. IEEE Transactions on Medical Imaging **41**(7), 1837–1848 (2022) 5, 9
26. Podobnik, G., Strojan, P., Peterlin, P., Ibragimov, B., Vrtovec, T.: Han-seg: The head and neck organ-at-risk ct and mr segmentation dataset. Medical Physics **50**(3), 1917–1927 (2023) 9, 10
27. Radl, L., Jin, Y., Pepe, A., Li, J., Gsaxner, C., Zhao, F.h., Egger, J.: Avt: Multicenter aortic vessel tree cta dataset collection with ground truth segmentation masks. Data in Brief **40**, 107801 (2022) 9
28. Rakic, M., Wong, H.E., Ortiz, J.J.G., Cimini, B.A., Guttag, J.V., Dalca, A.V.: Tyche: Stochastic in-context learning for medical image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11159–11173 (2024) 2, 9, 10, 11, 12
29. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations. vol. 2025, pp. 28085–28128 (2025) 1, 4
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 234–241 (2015) 4
31. Ruan, J., Li, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. ACM Transactions on Multimedia Computing, Communications and Applications (2024) 4
32. Tang, S., Yan, S., Qi, X., Gao, J., Ye, M., Zhang, J., Zhu, X.: Few-shot medical image segmentation with high-fidelity prototypes. Medical Image Analysis **100**, 103412 (2025) 5

33. Wang, H., Ni, D., Wang, Y.: Recursive deformable pyramid network for unsupervised medical image registration. IEEE Transactions on Medical Imaging **43**(6), 2229–2240 (2024) 5, 6, 10
34. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d: A vision foundation model for general-purpose segmentation on volumetric medical images. IEEE Transactions on Neural Networks and Learning Systems **36**(10), 17599–17612 (2025) 4
35. Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.H., Chen, Y.W., Tong, R.: Mixed transformer u-net for medical image segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 2390–2394 (2022) 4
36. Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A.: Multi-atlas segmentation with joint label fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(3), 611–623 (2012) 5
37. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Towards segmenting everything in context. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1130–1140 (2023) 2, 9, 10, 12
38. Wittmann, B., Wattenberg, Y., Amiranashvili, T., Shit, S., Menze, B.: vesselfm: A foundation model for universal 3d blood vessel segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20874–20884 (2025) 4, 7, 10, 12
39. Wong, H.E., Ortiz, J.J.G., Guttag, J., Dalca, A.V.: Multiverseg: Scalable interactive segmentation of biomedical imaging datasets with in-context guidance. In: IEEE/CVF International Conference on Computer Vision. pp. 20966–20980 (2025) 1, 4
40. Wong, H.E., Rakic, M., Guttag, J., Dalca, A.V.: Scribbleprompt: fast and flexible interactive segmentation for any biomedical image. In: European Conference on Computer Vision. pp. 207–229 (2024) 2, 4
41. Wu, J., Xu, M.: One-prompt to segment all medical images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11302–11312 (2024) 4
42. Xie, S., Zhang, L., Niu, Z., Ye, F., Zhong, Q., Xie, D., Chen, Y.W., Lin, L.: Eicseg: Universal medical image segmentation via explicit in-context learning. IEEE Transactions on Medical Imaging **45**(1), 53–69 (2025) 4
43. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 578–588 (2024) 4
44. Xu, Q., Zhu, L., Liu, X., Lin, G., Long, C., Li, Z., Zhao, R.: Unlocking the power of sam 2 for few-shot segmentation. In: International Conference on Machine Learning. pp. 69950–69966 (2025) 5
45. Zhang, X., Wu, Y., Angelini, E., Li, A., Guo, J., Rasmussen, J.M., O'Connor, T.G., Wadhwa, P.D., Jackowski, A.P., Li, H., et al.: Mapseg: Unified unsupervised domain adaptation for heterogeneous medical image segmentation based on 3d masked autoencoding and pseudo-labeling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5851–5862 (2024) 4
46. Zhang, Z., Yu, Y., Xue, Y.: Rethinking roi strategy in interactive 3d segmentation for medical images. In: CVPR 2025 Challenge on Foundation Models for 3D Biomedical Image Segmentation. pp. 208–223 (2025) 1
47. Zhang, Z., Yu, Y., Xue, Y.: Unlocking 2d promptable foundation models for 3d vessel segmentation by automatic prompt generation. In: Medical Imaging with Deep Learning. pp. 3764–3778 (2026) 1

48. Zhao, J., Yang, F., Li, X., Jiao, Z., Zhai, Q., Li, X., Wu, D., Fu, H., Cheng, H.: Segmic: A universal model for medical image segmentation through in-context learning. Pattern Recognition **171**, 112179 (2025) 4
49. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE Transactions on Image Processing **32**, 4036–4045 (2023) 4
50. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging **39**(6), 1856–1867 (2019) 4
51. Zhu, Y., Zhang, H.: Maup: Training-free multi-center adaptive uncertainty-aware prompting for cross-domain few-shot medical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 326–336 (2025) 5