Medverse: A Universal Model for Full-Resolution 3D Medical Image Segmentation, Transformation and Enhancement

Jiesi Hu[1,2], Jianfeng Cao[1*], Yanwu Yang[3,4], Chenfei Ye[1], Yixuan Zhang[1], Hanyang Peng[2*], Ting Ma[1,2*]

[1]Harbin Institute of Technology at Shenzhen, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]University Hospital Tübingen, Tübingen, Germany
[4]German Center for Mental Health, Germany
405323011@qq.com, caojianfeng@hit.edu.cn, philoso_phy0922@163.com, tma@hit.edu.cn

Abstract

In-context learning (ICL) offers a promising paradigm for universal medical image analysis, enabling models to perform diverse image processing tasks without retraining. However, current ICL models for medical imaging remain limited in two critical aspects: they cannot simultaneously achieve high-fidelity predictions and global anatomical understanding, and there is no unified model trained across diverse medical imaging tasks (e.g., segmentation and enhancement) and anatomical regions. As a result, the full potential of ICL in medical imaging remains underexplored. Thus, we present **Medverse**, a universal ICL model for 3D medical imaging, trained on 22 datasets covering diverse tasks in universal image segmentation, transformation, and enhancement across multiple organs, imaging modalities, and clinical centers. Medverse employs a next-scale autoregressive in-context learning framework that progressively refines predictions from coarse to fine, generating consistent, full-resolution volumetric outputs and enabling multi-scale anatomical awareness. We further propose a blockwise cross-attention module that facilitates long-range interactions between context and target inputs while preserving computational efficiency through spatial sparsity. Medverse is extensively evaluated on a broad collection of held-out datasets covering previously unseen clinical centers, organs, species, and imaging modalities. Results demonstrate that Medverse substantially outperforms existing ICL baselines and establishes a novel paradigm for in-context learning.

Code — https://github.com/jiesihu/Medverse
Extended version — https://arxiv.org/pdf/2509.09232

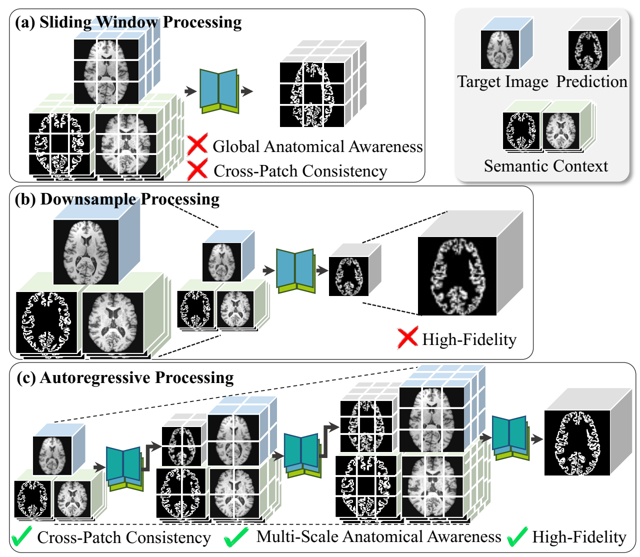

Figure 1: Illustration of different strategies for processing 3D medical images using ICL models. Due to the high resolution of volumetric data, direct end-to-end processing is often infeasible. The proposed next-scale autoregressive ICL framework progressively refines predictions from coarse to fine, producing full-resolution outputs.

Introduction

Universal medical imaging models trained on large-scale, diverse datasets represent a significant step toward clinical AI deployment, exhibiting strong generalization across clinical centers and medical tasks (Butoi et al. 2023; Yang et al. 2024, 2025).

Initially introduced in natural language processing (Brown et al. 2020), in-context learning (ICL) has recently emerged as a unified paradigm in medical imaging (Butoi et al. 2023; Hu et al. 2024, 2025b; Gao et al. 2025). It conditions on a set of image–label pairs from other subjects, known as the context, to convey task information. By varying these context examples, a single model can perform diverse tasks such as segmentation, transformation, and enhancement, even including tasks that were not encountered during training (Czolbe and Dalca 2023). Unlike traditional adaptation approaches, ICL adapts to distribution shifts and novel tasks without retraining (Chan et al. 2022; Reddy 2023), making it particularly suited for heterogeneous medical imaging scenarios.

SegGPT (Wang et al. 2023b) and Painter (Wang et al. 2023a) have demonstrated the effectiveness of ICL in nat-

*Corresponding authors.

ural image domains. However, their performance is suboptimal when applied to medical images. Models such as UniverSeg (Butoi et al. 2023) and One-Prompt (Wu and Xu 2024), trained on medical imaging data, have shown promising results on 2D medical image segmentation tasks. Neuralizer (Czolbe and Dalca 2023) further extends ICL to a wider range of 2D tasks, including denoising and skull stripping. However, when applied to 3D data, they suffer from the loss of topological correlations and volumetric context. More recently, Neuroverse3D (Hu et al. 2025a) has extended ICL to 3D neuroimaging by introducing adaptive context processing and architectural designs addressing the computational challenges of volumetric data.

However, a major challenge is that current models remain constrained to operate at predetermined low spatial resolutions, which restricts their ability to capture fine-grained anatomical details and preserve the fidelity of output images. Attempts to increase output resolution via sliding-window strategies face the challenge of disrupting the anatomical continuity of context and target images, which impairs accurate understanding and the transmission of task-specific guidance from the context. This creates a dilemma for the development and practical use of ICL models in medical imaging. Another pressing issue is the absence of a unified 3D ICL model in medical imaging that is jointly trained across multiple organs and diverse types of image processing tasks, leaving the full potential of ICL in this domain underexplored.

To overcome the aforementioned issues, we present Medverse, a universal 3D medical image model trained on a diverse collection of datasets spanning multiple organs, imaging modalities, clinical centers, and task types, including universal segmentation, transformation, and enhancement. We introduce a **Next-Scale Autoregressive ICL framework (NA-ICL)** that adopts a coarse-to-fine prediction strategy and supports arbitrary input resolutions (Figure 1). By integrating multi-scale anatomical cues from both context and target images, NA-ICL effectively balances global semantics with local details, improving prediction accuracy and producing consistent, full-resolution outputs while mitigating sliding-window artifacts. To enhance context perception under complex real-world conditions, we introduce a **Blockwise Cross-Attention Module (BAM)**, which enables long-range context-target interactions and alleviates spatial misalignment, while preserving computational and memory efficiency through spatial sparsity.

Our contributions are summarized as follows:

- We introduce NA-ICL, a next-scale autoregressive in-context learning framework that enables the model to perform coarse-to-fine prediction. This allows the model to leverage global semantics and local details, producing high-fidelity and spatially consistent, full-resolution outputs.
- We introduce BAM, which enables long-range context-target interactions while maintaining minimal memory and computational overhead through spatial sparsity.
- Trained on a diverse collection of datasets, Medverse is extensively evaluated on held-out data spanning unseen centers, organs, species, and modalities. It achieves strong generalization and consistently outperforms existing in-context learning baselines across segmentation, transformation, and enhancement tasks.

Method

Under the ICL paradigm, Medverse achieves universality by learning task-specific guidance from context examples in the form of image-label pairs. It is capable of performing diverse tasks by leveraging the rich information embedded in these contexts.

Next-Scale Autoregressive ICL

Unlike conventional ICL models that utilize only context from other subjects, our NA-ICL framework provides the model with both image-label pairs from other subjects and the model's coarse predictions at lower resolutions, inspired by (Tian et al. 2024). This architectural design enables progressive processing from coarse to fine scales, offering benefits including the integration of multi-scale anatomical information, improved consistency across patches, and high-fidelity outputs.

Network Architecture. As shown in Figure 2 (a), the proposed model comprises three 3D U-Net branches that, from top to bottom, process the autoregressive context, the target image, and the semantic context. The target image branch interacts with both context branches through fusion modules to facilitate the context-target feature interaction.

The autoregressive context comprises the target image and its associated prediction from a lower-resolution step, and is treated as a specialized type of context. The semantic context consists of a set of image-label pairs drawn from other subjects, providing conventional semantic guidance for the task. To promote parameter efficiency and shared contextual understanding, the autoregressive context branch shares weights with the semantic context branch. To differentiate the features from the autoregressive context, we introduce a learnable autoregressive embedding $\boldsymbol{e} \in \mathbb{R}^C$, which is added to the output of the first layer of the autoregressive context branch as follows:

$$\boldsymbol{f}^{'}_{1,\text{AR}} = \boldsymbol{f}_{1,\text{AR}} + \text{Expand}(\boldsymbol{e}), \tag{1}$$

where $\boldsymbol{f}_{1,\text{AR}} \in \mathbb{R}^{C\times H\times W\times D}$ denotes the feature map from the first layer and $\text{Expand}(\cdot)$ expands the embedding across spatial dimensions.

Let the semantic context be defined as $S^{(t)}$, which contains a number of image-label examples. t refers to the corresponding autoregressive step, which will be detailed later. Given the target image $\boldsymbol{x}^{(t)}$ and the autoregressive context from the previous step, $\mathcal{A}^{(t-1)} = (\boldsymbol{x}^{(t-1)}, \hat{\boldsymbol{y}}^{(t-1)})$, the model prediction at step t is computed as

$$\hat{\boldsymbol{y}}^{(t)} = F(\boldsymbol{x}^{(t)}, S^{(t)}, \mathcal{A}^{(t-1)}), \tag{2}$$

where $F(\cdot)$ denotes the prediction function of Medverse. To efficiently handle a large number of semantic context examples, we incorporate the adaptive parallel-sequential context processing module introduced in (Hu et al. 2025a), which

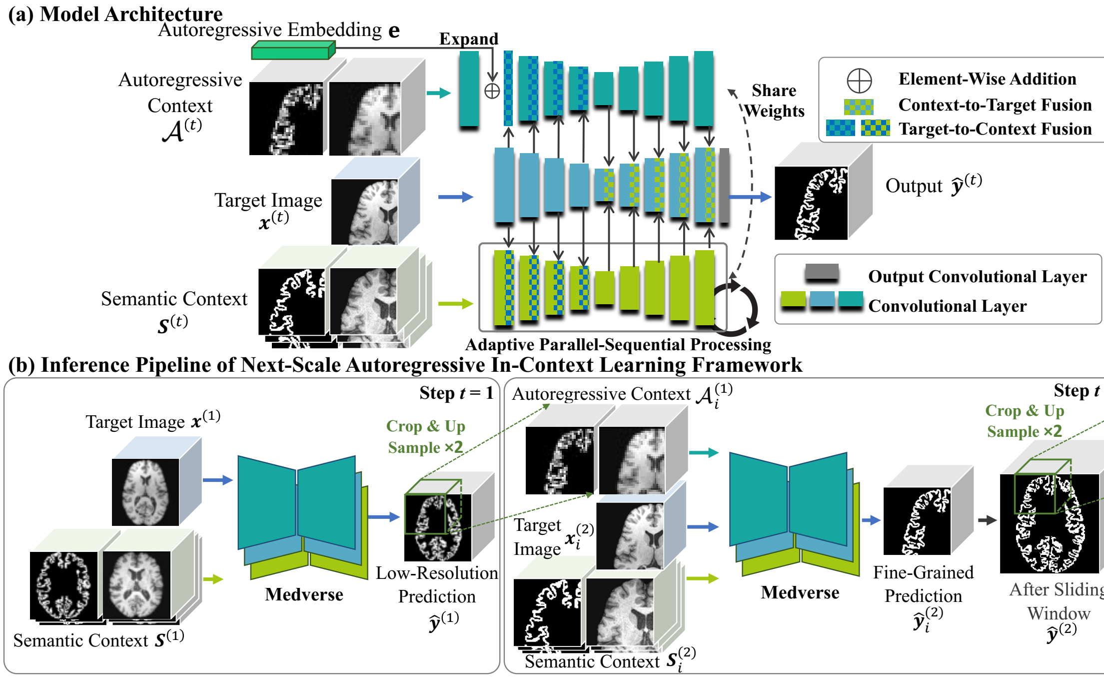

Figure 2: Illustration of our model architecture and the inference pipeline of the next-scale autoregressive in-context learning framework.

enables scalable context processing with minimal memory requirements.

Autoregressive Inference Pipeline. Given an input volume of size (H, W, D), the number of autoregressive steps is

$$T = \lceil \log_2(\frac{\max\{H, W, D\}}{I}) \rceil + 1, \tag{3}$$

where $I \times I \times I$ denotes the model's input patch size. For step $t \in \{1, \ldots, T\}$, we downsample the input images to

$$\left(H^{(t)}, W^{(t)}, D^{(t)}\right) = \lceil \frac{H}{2^{T-t}} \rceil, \lceil \frac{W}{2^{T-t}} \rceil, \lceil \frac{D}{2^{T-t}} \rceil, \tag{4}$$

so that the resolution is doubled from level t to $t+1$. We assume that the context and target images share the same spatial dimensions and undergo identical resizing operations. After the final step $t = T$, the model outputs a prediction at the original image resolution.

Figure 2 (b) illustrates the autoregressive inference pipeline. At step $t = 1$, the model predicts $\hat{\boldsymbol{y}}^{(1)} = F(\boldsymbol{x}^{(1)}, S^{(1)}, \mathcal{A}^{(0)})$ on the low-resolution input image, capturing global anatomical context, with $\mathcal{A}^{(0)}$ initialized as an empty set. The autoregressive context $\mathcal{A}^{(1)} = (\boldsymbol{x}^{(1)}, \hat{\boldsymbol{y}}^{(1)})$ is then passed to step $t = 2$. At $t = 2$, the increased resolution requires the target image $\boldsymbol{x}^{(2)}$ to be split into $I \times I \times I$ patches, which are processed with a sliding window to capture finer anatomical detail. The prediction for each patch $\boldsymbol{x}_i^{(2)}$ is then given by $\hat{\boldsymbol{y}}_i^{(2)} = F(\boldsymbol{x}_i^{(2)}, S_i^{(2)}, \mathcal{A}_i^{(1)})$. $S_i^{(2)}$ denotes the semantic context cropped at the same spatial location as $\boldsymbol{x}_i^{(2)}$ to ensure patch-wise alignment. To propagate global information from coarser scales, the corresponding region in $\mathcal{A}^{(1)}$ of size $\frac{I}{2} \times \frac{I}{2} \times \frac{I}{2}$ is extracted, upsampled to $I \times I \times I$, and used as $\mathcal{A}_i^{(1)}$. The same procedure is repeated for $t \geq 2$, with the number of sliding windows increasing as the resolution increases. A step-by-step pseudocode is provided in the supplementary material.

Context-Target Fusion Module

Feature interaction between different branches in the network is achieved through fusion modules, enabling the target branch to effectively incorporate contextual information. As shown in Figure 2 (a), we perform target-to-context fusion within the encoder, where features from the target branch are passed to the context branch, and context-to-target fusion within the decoder, where features from the context branch are passed back to the target branch. The use of adaptive parallel-sequential context processing necessitates this order of interaction.

The details of the fusion module are illustrated in Figure 3. The target-to-context module employs the context feature map as the query, with the target feature map serving as the source of keys and values. BAM transfers target information to the context, after which the output is concatenated with the original context, compressed by a $1 \times 1 \times 1$ convolution, and added back through a residual shortcut to produce a target-aware context representation.

The context-to-target fusion module concatenates semantic and autoregressive context features, using the combined representation as the source of keys and values, with the target features serving as the query. BAM injects both semantic and autoregressive contextual guidance into the target representation. The resulting features are then similarly passed

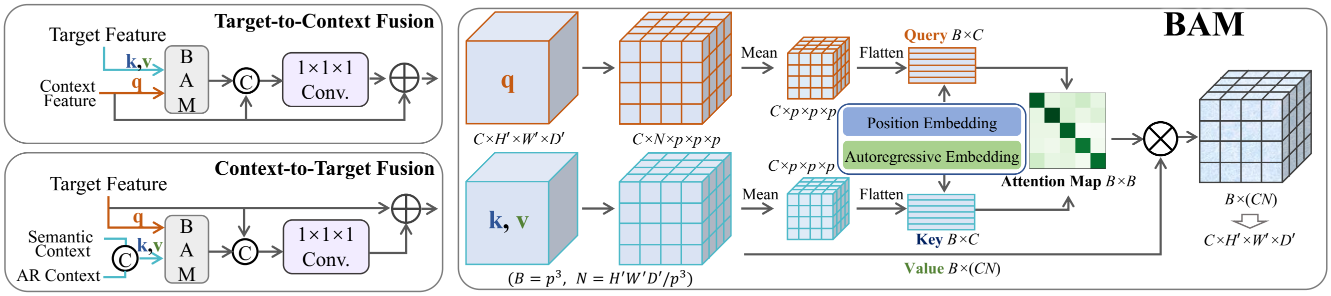

Figure 3: Illustration of fusion modules and the blockwise cross-attention module.

through concatenation, convolution, and a residual shortcut to produce a context-aware target representation.

Blockwise Cross-Attention Module. Since there is not necessarily strict spatial alignment between the context and target objects, long-range interactions are crucial for effective feature fusion. We propose BAM, which enables efficient long-range feature aggregation with spatial sparsity.

We denote the query feature map by $\boldsymbol{X}_q \in \mathbb{R}^{C\times H'\times W'\times D'}$, and the key–value map by $\boldsymbol{X}_{kv} \in \mathbb{R}^{C\times H'\times W'\times D'}$. The volume is partitioned into $p\times p\times p$ non-overlapping blocks, so that each block covers $h=\frac{H'}{p}$, $w=\frac{W'}{p}$, and $d=\frac{D'}{p}$ voxels. Let $N=hwd$ denote the number of voxels per block. After partitioning, we obtain the blockwise features $\boldsymbol{X}_q^{\text{Blockwise}}, \boldsymbol{X}_{kv}^{\text{Blockwise}} \in \mathbb{R}^{C\times N\times p\times p\times p}$. The query and key representations for each block are then computed by mean pooling over the N voxels within the block:

$$\boldsymbol{Q}' = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_q^{\text{Blockwise}}[:,n,:,:,:] \in \mathbb{R}^{C\times p\times p\times p}, \quad (5)$$

$$\boldsymbol{K}' = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_{kv}^{\text{Blockwise}}[:,n,:,:,:] \in \mathbb{R}^{C\times p\times p\times p}. \quad (6)$$

To match the input format required by the attention mechanism, we transpose and reshape $\boldsymbol{Q}'$ and $\boldsymbol{K}'$ to obtain $\boldsymbol{Q}, \boldsymbol{K} \in \mathbb{R}^{B\times C}$, where $B=p^3$ denotes the total number of blocks. Then, the queries and keys are linearly projected into a m-dimensional space via learnable matrices $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{C\times m}$. To incorporate spatial priors, we add a 3D sine–cosine positional embedding $\boldsymbol{P} \in \mathbb{R}^{B\times m}$ to both branches. In addition, for blocks originating from the autoregressive context, we further add a learnable autoregressive embedding $\boldsymbol{R} \in \mathbb{R}^{m}$, which acts as a marker for features from the autoregressive branch and is broadcast across all blocks to shape $B\times m$. The resulting transformed features are given by

$$\widehat{\boldsymbol{Q}} = \boldsymbol{Q}\boldsymbol{W}_Q + \boldsymbol{P} + \mathbf{1}_{\text{AR}}\cdot\boldsymbol{R}, \quad (7)$$

$$\widehat{\boldsymbol{K}} = \boldsymbol{K}\boldsymbol{W}_K + \boldsymbol{P} + \mathbf{1}_{\text{AR}}\cdot\boldsymbol{R}, \quad (8)$$

where $\mathbf{1}_{\text{AR}}$ is an indicator that equals 1 if the block comes from the autoregressive context and 0 otherwise. The block-level attention weights are as follows:

$$\boldsymbol{A} = \text{Softmax}\Big(\frac{\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{K}}^{\top}}{\sqrt{m}}\Big) \in \mathbb{R}^{B\times B}. \quad (9)$$

To preserve spatial granularity, the value features are obtained directly from the unpooled input as $\boldsymbol{V}' = \boldsymbol{X}_{kv}^{\text{Blockwise}} \in \mathbb{R}^{C\times N\times p\times p\times p}$. They are then transposed and reshaped to match the input format for attention dot-product computation, yielding $\boldsymbol{V} \in \mathbb{R}^{B\times(CN)}$. The value matrix $\boldsymbol{V}$ is multiplied by the attention weights to obtain

$$\bar{\boldsymbol{Y}} = \boldsymbol{A}\,\boldsymbol{V} \in \mathbb{R}^{B\times(CN)}.$$

The result is then reshaped back to the original spatial layout, producing the final BAM output:

$$\widehat{\boldsymbol{Y}} = \text{Reshape}\big(\bar{\boldsymbol{Y}}\big) \in \mathbb{R}^{C\times H'\times W'\times D'}.$$

The computational complexity is thereby reduced to $\mathcal{O}(B^2)$, in contrast to the original $\mathcal{O}\big((H'W'D')^2\big)$, while still preserving long-range interactions and retaining fine-grained value information. $B=p^3$ is a user-defined constant that remains fixed across U-Net stages, further ensuring stable computational cost. Compared to the direct concatenation-based fusion used in (Butoi et al. 2023; Czolbe and Dalca 2023; Hu et al. 2025a), this cross-attention fusion approach is better suited for handling spatial misalignment between the context and target objects.

Loss Function

For fair comparison, we adopt the same loss functions as those proposed in (Hu et al. 2025a). Specifically, a modified $\mathcal{L}_{\text{smooth}-L_1}$ loss is used for segmentation tasks, while for image enhancement and transformation tasks, the $\mathcal{L}_{\text{smooth}-L_1}$ loss is applied to both image intensity and intensity differences.

Experiments

Data and Tasks

Datasets. To ensure robust cross-center generalization and data diversity, we curated a collection of **27 publicly available datasets** spanning multiple imaging modalities and acquisition centers, comprising a total of **40,362 3D scans**. The dataset encompasses widely used medical imaging modalities, including T1, T2, FLAIR, MRA, DWI, ADC, PD, and

Methods	Fine-Tuning Free	*Unseen Center*						*Unseen Organ*			*Unseen Species*	*Unseen Modality*	Average
		Cerebral Cortex	Hippocampus	Thalamus	Liver	Spleen	Kidney Left	Maxillary Sinus	Nasal Cavity	Nasal Pharynx	Mice Lung	PET Lateral Ventricle	
Fully Supervised Task-Specific Models (Upper Bound)													
nnUNet	✗	90.30	90.99	93.89	98.46	96.60	96.06	94.07	91.63	94.63	94.49	84.26	93.22
3D-Unet	✗	88.55	89.73	92.88	96.41	96.52	91.10	90.13	89.02	92.64	94.21	82.35	91.23
Swin-UNETR	✗	89.78	89.38	92.92	96.49	94.45	94.88	94.79	89.08	93.65	93.21	82.11	91.88
Few-Shot Task-Specific Models													
3D-Unet	✗	87.90	86.66	90.56	94.95	81.74	81.29	86.77	86.99	90.05	91.89	75.95	86.80
Swin-UNETR	✗	87.62	86.30	91.15	94.66	88.64	87.82	87.99	84.96	89.46	91.40	74.40	87.67
ICL Models													
SegGPT	✓	45.38	28.41	19.56	68.07	39.02	36.15	46.35	52.79	37.25	43.30	42.22	41.68
Neuralizer	✓	69.20	57.49	45.11	73.54	52.12	62.71	75.77	64.79	73.65	70.48	51.83	63.34
UniverSeg	✓	68.79	59.90	47.57	81.10	57.79	56.76	80.12	75.78	72.64	65.77	48.90	65.01
Neuroverse3D	✓	85.69	**83.98**	**89.98**	93.67	82.66	75.75	78.08	74.66	**87.23**	80.55	59.83	81.10
Medverse	✓	**87.30**	82.12	87.65	**95.90**	**91.05**	**95.31**	**92.63**	**78.15**	87.13	**92.21**	**70.48**	**87.27**

Table 1: Segmentation comparison across unseen centers, organs, species, and modalities in terms of Dice scores (%). The context set and the training set for few-shot models are both of size 4.

Methods	Fine-Tuning Free	*Transformation*		*Enhancement*				Average
		Skull Stripping	Modality Transform	Bias Removal	Gaussian Noise	Salt Pepper	Inpainting	
Fully Supervised Task-Specific Models (Upper Bound)								
3D U-Net	✗	26.18	32.97	30.09	31.86	38.22	34.57	32.31
Swin-UNETR	✗	25.23	29.75	29.60	29.64	33.29	31.09	29.77
Few-Shot Task-Specific Models								
3D U-Net	✗	22.78	28.74	27.34	29.22	34.65	32.23	29.16
Swin-UNETR	✗	23.64	27.39	27.26	25.08	29.41	25.50	26.38
ICL Models								
Painter	✓	9.15	13.43	15.60	14.86	13.90	12.38	13.22
Neuralizer	✓	21.01	24.71	26.13	13.14	25.26	23.75	22.34
Neuroverse3D	✓	26.31	24.37	28.42	25.71	30.65	27.01	27.08
Medverse	✓	**26.36**	**24.80**	**28.48**	**26.08**	**32.57**	**31.51**	**28.30**

Table 2: Performance comparison across image transformation and enhancement tasks with a context size of 4 in terms of PSNR. Enhancement tasks are the average over 5 held-out datasets.

CT, as well as commonly studied anatomical regions such as the **brain**, **abdomen**, **prostate**, and **lung**. 22 datasets (Yang et al. 2023; CAS 2023; Hernandez Petzsche et al. 2022; Liew et al. 2022; IXI 2015; Hoopes et al. 2022; Marcus et al. 2007; Flanders et al. 2020; ADHD 2011; Jack Jr et al. 2008; Alexander et al. 2017; Holmes et al. 2015; Gera et al. 2023; Nugent et al. 2022; Sudlow et al. 2015; Menze et al. 2014; Litjens et al. 2014; Kuijf et al. 2019; Zeng et al. 2023; Ji et al. 2022; Luo et al. 2024; Wasserthal et al. 2023; Antonelli et al. 2022) were used for training and validation with a random 9:1 split. The remaining five datasets were reserved as held-out sets to evaluate generalization to unseen distributions. These held-out datasets include brain and abdomen scans from previously unseen centers (Marek et al. 2011; Ma et al. 2024), the nasal cavity as an unseen anatomical target (Zhang et al. 2024), mice as an unseen species (Rosenhain et al. 2018), and PET as an unseen imaging modality (Jack Jr et al. 2008). Each held-out dataset was split in a 5:5 ratio into a meta context set used for semantic context selection and a test set.

Data Preprocessing. To improve the model's generalizability across acquisition centers, no spatial resampling was applied. Image intensities were normalized to the range from 0 to 1 using the 0.02 and 0.98 percentile values as lower and upper bounds, respectively. Segmentation masks were binarized by assigning zero to the background and one to the foreground.

Tasks. We trained our model on three categories of tasks, encompassing segmentation, transformation, and enhancement objectives. The segmentation tasks include anatomical structure delineation across various organs, as well as tumor and vessel segmentation. The transformation tasks include arbitrary modality-to-modality transformation and brain extraction for skull stripping. The enhancement tasks involve improving image quality through bias field correction, image inpainting, and the removal of Gaussian and salt-and-pepper noise.

Further details regarding datasets, task definitions, training procedures, and evaluation protocols are provided in the supplemental material.

Compared Models

Medverse. Both the target and context branches are implemented using a five-stage 3D U-Net architecture (Çiçek et al. 2016). Each branch begins with 32 channels in the first stage, with the number of channels doubling at each subsequent stage. The model operates on input patches of size $128 \times 128 \times 128$. In BAM, p is set to 4, and m is set to 66.

Task-Specific Models. We compared Medverse against task-specific 3D models. Each task-specific model was trained directly on the meta context set of the held-out datasets, thereby mitigating domain shift concerns. This setting reflects the conventional practice in clinical environments, where models are trained specifically for a given task. Two backbone architectures were evaluated: a 3D U-Net, which shares the same architecture and channel configuration as the Medverse backbone, and the Swin-

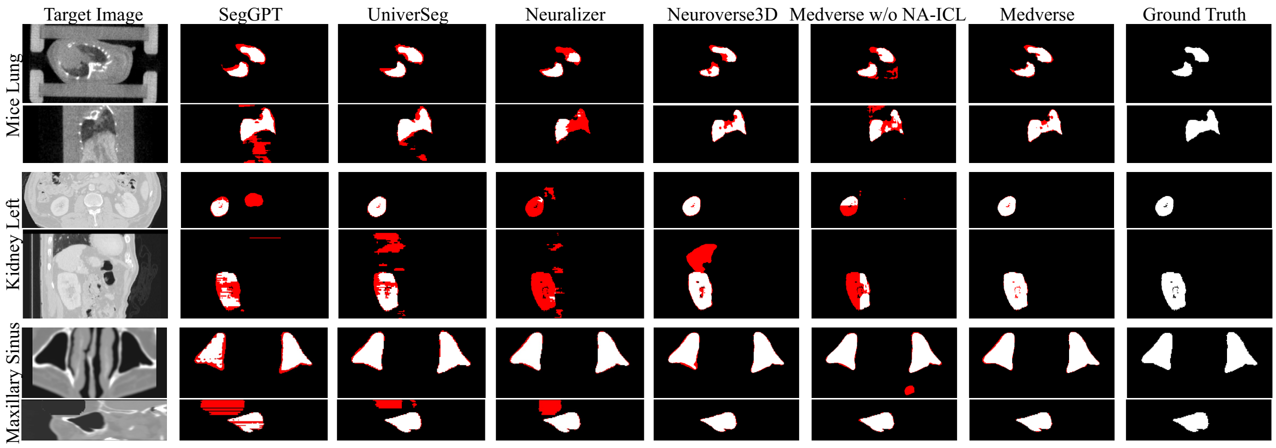

Figure 4: Qualitative results of ICL models on segmentation tasks. For each 3D segmentation target, results are shown from two views. Red regions indicate segmentation errors. The 2D models take the slice corresponding to the first view as input. Medverse w/o NA-ICL denotes a variant of Medverse without autoregressive processing.

UNETR (Hatamizadeh et al. 2021). We assessed performance under both few-shot and fully supervised learning scenarios.

Other ICL Models. We compared our method with several state-of-the-art 2D ICL models, including Painter (Wang et al. 2023a), SegGPT (Wang et al. 2023b), UniverSeg (Butoi et al. 2023), and Neuralizer (Czolbe and Dalca 2023). To apply them to 3D data, each 3D target image was sliced into 2D planes, and 2D predictions were reassembled into 3D volumes for evaluation. Slices covering the target region were randomly selected to form the 2D context set. The number of context slices followed the best settings reported in their papers: 1 for Painter, 8 for SegGPT, 32 for Neuralizer, and 64 for UniverSeg. We also included Neuroverse3D (Hu et al. 2025a), an ICL model designed for 3D neuroimaging. Because it only supports inputs of size $128 \times 128 \times 128$, we resized images before inference and scaled the outputs back to the original resolution. All models were tested using their publicly available pretrained weights.

Results on Held-Out Dataset

Quantitative Comparison on Segmentation Tasks. Table 1 reports the segmentation performance of the models. Medverse achieves the highest performance among all ICL methods, surpassing the second-best method, Neuroverse3D, by about 6 points in the average Dice score. This improvement can be attributed to Medverse's broader training data coverage and its superior ability to capture fine details enabled by the NA-ICL framework. 2D models, such as UniverSeg and SegGPT, exhibit decreased performance on 3D volumes, primarily due to their tendency to produce false positives in slices that do not contain the target anatomy. When compared to few-shot task-specific models, Medverse achieves similar levels of performance while offering the notable advantage of requiring no task-specific fine-tuning. These results highlight the strong potential and practical value of our method. We further evaluate our model across more classes on the validation set, with results provided in the Supplementary Table 4.

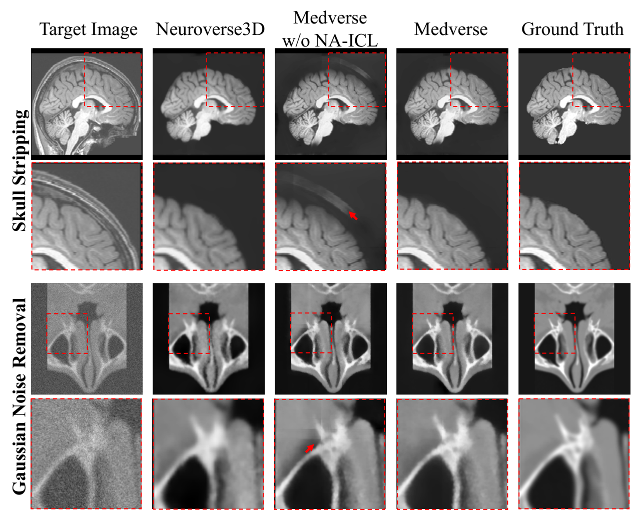

Figure 5: Qualitative results of 3D ICL models. Medverse w/o NA-ICL denotes the variant of Medverse without autoregressive context. The second row for each task presents zoomed-in views highlighting differences in generated details. Neuroverse3D produces outputs with limited resolution. Red arrows indicate artifacts.

Quantitative Comparison on Transformation and Enhancement Tasks. Table 2 presents the performance of different models on transformation and enhancement tasks. Medverse significantly outperforms other ICL models on tasks such as inpainting and denoising. Although Medverse achieves only a modest improvement of 1.22 PSNR over the second-best Neuroverse3D on average, it preserves the full resolution of the input image, whereas Neuroverse3D limits the resolution of its outputs. Despite being the best among ICL models, Medverse still lags behind the fully supervised

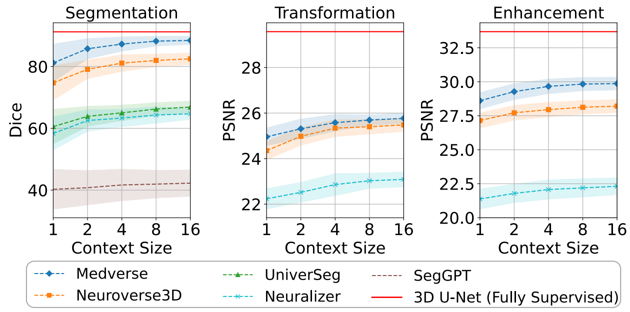

Figure 6: Performance comparison of ICL models under varying 3D context sizes.

BAM	NA-ICL	Seg. (Dice ↑)	Tran. (PSNR ↑)	Enh. (PSNR ↑)
		76.50	24.53	27.45
✓		78.90	25.38	27.93
✓	✓	**87.27**	**25.58**	**29.66**

Table 3: Ablation study on the BAM module and NA-ICL framework. BAM is replaced with plain feature concatenation when removed, and NA-ICL is replaced with sliding window processing.

3D U-Net and shows a performance gap compared to the few-shot 3D U-Net.

Qualitative Results on Segmentation Tasks. Figure 4 presents the segmentation results of different ICL models. The 2D models perform reasonably well on individual slices that contain the target structure but exhibit discontinuities across adjacent slices, and SegGPT and UniverSeg tend to produce false positives in slices without the target. Neuroverse3D alleviates the limitations of 2D processing but struggles to generalize to unseen organ. Medverse w/o NA-ICL refers to applying a sliding window directly, without autoregressive context. In this case, the model lacks access to sufficient global context, leading to unstable predictions. In contrast, Medverse yields the most accurate and fine-grained segmentation results.

Qualitative Results on Transformation and Enhancement Tasks. Figure 5 illustrates the performance of 3D ICL models on skull stripping and Gaussian noise removal tasks. As shown, Neuroverse3D fails to preserve image resolution, resulting in a loss of fidelity, whereas Medverse effectively retains detailed structures. Without the NA-ICL framework, the model struggles to accurately remove the skull, as having access only to local information increases task difficulty. In the Gaussian noise removal task, stitching artifacts appear as visible lines, which are substantially mitigated when the NA-ICL framework is employed.

Additional qualitative results on segmentation, transformation, and enhancement can be found in the supplementary material.

Effect of Context Size. Figure 6 illustrates how model performance varies with context size. Consistent with prior work, all models improve as context increases. For Medverse, the Dice score on segmentation rises by 7.28 points when context grows from 1 to 16, while PSNR increases by 0.81 and 1.26 for transformation and enhancement tasks. The larger segmentation gain likely reflects sparser labels, whereas transformation and enhancement tasks already provide richer semantic cues, yielding smaller marginal gains. Medverse consistently outperforms other ICL models. In transformation tasks, Neuroverse3D performs similarly due to its neuroimaging training domain but slightly lags because of lower output resolution.

Computational Efficiency. BAM processes only 4^3 tokens, adding just 2.01×10^{-1} GFLOPs and 9.18×10^{-4} memory overhead per context pair. In contrast, standard cross-attention with 128^3 tokens incurs 2.36×10^6 GFLOPs and 2.55 memory overhead, making it computationally impractical. Additionally, our model maintains a fixed inference memory usage of 9.14 GB regardless of context size, enabled by adaptive parallel-sequential context processing. Inference on a $128 \times 128 \times 128$ patch with 8 context samples and 1 autoregressive context takes only 1.16 seconds, whereas models like Painter and SegGPT exceed 30 seconds. These properties make our model both efficient and practical for deployment. Full details are provided in the supplemental material.

Ablation Analysis of Key Components. Table 3 shows ablations on the BAM module and the NA-ICL framework. Removing NA-ICL, which replaces coarse-to-fine prediction with direct sliding-window inference, leads to a notable drop in segmentation performance. This is due to the limited perspective field of the target image and the loss of global context guidance. For transformation and enhancement tasks, NA-ICL also improves performance, with one contributing factor being its ability to enhance consistency across sliding-window patches. When the BAM module is removed, feature fusion is instead performed via simple concatenation as in (Hu et al. 2025a). The BAM module leads to overall performance improvements due to mitigating the adverse effects caused by spatial misalignment between the context and target images.

Conclusion

We present Medverse, a universal in-context learning model for 3D medical image analysis that handles segmentation, transformation, and enhancement across diverse organs, modalities, and centers. Trained on large-scale heterogeneous data, Medverse generalizes to unseen domains without task-specific fine-tuning. To ensure high-fidelity outputs, we propose a next-scale autoregressive framework for multi-scale anatomical awareness and a blockwise Cross-Attention Module for efficient long-range context–target interaction. Extensive experiments show that Medverse surpasses existing ICL models and sets a new paradigm for universal 3D medical image processing. Despite its limited parameters and training data, Medverse demonstrates strong potential for further scaling and broader applications in medical imaging.

Acknowledgements

This work was supported in part by grants from the National Natural Science Foundation of P.R. China (62276081), the Basic Foundation of Shenzhen Science and Technology National Key Program (CJGJZD20230724093959002, JCYJ20250604145427037), the Guangdong Basic and Applied Basic Research Foundation (2023A1515010792, 2023B151512006, 2025A1515010212), the Shenzhen Science and Technology Program (GXWD20231129121139001, JCYJ20240813110522029), the Major Key Project of PCL (PCL2025A09), and the Harbin Institute of Technology, Shenzhen (YG4500100125). Yanwu acknowledges the affiliation with the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

ADHD 2011. 2011. ADHD-200. https://fcon_1000.projects.nitrc.org/indi/adhd200/.

Alexander, L. M.; Escalera, J.; Ai, L.; Andreotti, C.; Febre, K.; Mangone, A.; Vega-Potler, N.; Langer, N.; Alexander, A.; Kovacs, M.; et al. 2017. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific data*, 4(1): 1–26.

Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R. M.; et al. 2022. The medical segmentation decathlon. *Nature communications*, 13(1): 4128.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Butoi, V. I.; Ortiz, J. J. G.; Ma, T.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2023. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21438–21451.

CAS 2023. 2023. Cerebral artery segmentation Challenge (CAS) 2023. https://codalab.lisn.upsaclay.fr/competitions/9804.

Chan, S.; Santoro, A.; Lampinen, A.; Wang, J.; Singh, A.; Richemond, P.; McClelland, J.; and Hill, F. 2022. Data distributional properties drive emergent in-context learning in transformers. *Advances in neural information processing systems*, 35: 18878–18891.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 424–432. Springer.

Czolbe, S.; and Dalca, A. V. 2023. Neuralizer: General Neuroimage Analysis without Re-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6217–6230.

Flanders, A. E.; Prevedello, L. M.; Shih, G.; Halabi, S. S.; Kalpathy-Cramer, J.; Ball, R.; Mongan, J. T.; Stein, A.; Kitamura, F. C.; Lungren, M. P.; et al. 2020. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3): e190211.

Gao, Y.; Liu, D.; Li, Z.; Li, Y.; Chen, D.; Zhou, M.; and Metaxas, D. N. 2025. Show and Segment: Universal Medical Image Segmentation via In-Context Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20830–20840.

Gera, R.; Or, M. B.; Tavor, I.; Roll, D.; Cockburn, J.; Barak, S.; Tricomi, E.; O'Doherty, J. P.; and Schonberg, T. 2023. Characterizing habit learning in the human brain at the individual and group levels: A multi-modal MRI study. *NeuroImage*, 272: 120002.

Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H. R.; and Xu, D. 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, 272–284. Springer.

Hernandez Petzsche, M. R.; de la Rosa, E.; Hanning, U.; Wiest, R.; Valenzuela, W.; Reyes, M.; Meyer, M.; Liew, S.-L.; Kofler, F.; Ezhov, I.; et al. 2022. ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1): 762.

Holmes, A. J.; Hollinshead, M. O.; O'keefe, T. M.; Petrov, V. I.; Fariello, G. R.; Wald, L. L.; Fischl, B.; Rosen, B. R.; Mair, R. W.; Roffman, J. L.; et al. 2015. Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. *Scientific data*, 2(1): 1–16.

Hoopes, A.; Hoffmann, M.; Greve, D. N.; Fischl, B.; Guttag, J.; and Dalca, A. V. 2022. Learning the effect of registration hyperparameters with hypermorph. *The journal of machine learning for biomedical imaging*, 1.

Hu, J.; Peng, H.; Yang, Y.; Guo, X.; Shang, Y.; Shi, P.; Ye, C.; and Ma, T. 2025a. Building 3D In-Context Learning Universal Model in Neuroimaging. *arXiv preprint arXiv:2503.02410*.

Hu, J.; Shang, Y.; Yang, Y.; Guo, X.; Peng, H.; and Ma, T. 2024. Icl-sam: Synergizing in-context learning model and sam in medical image segmentation. *Medical Imaging with Deep Learning*, 641–656.

Hu, J.; Yang, Y.; Ye, Z.; Zhou, J.; Cao, J.; Peng, H.; and Ma, T. 2025b. Efficient Universal Models for Medical Image Segmentation via Weakly Supervised In-Context Learning. *arXiv preprint arXiv:2510.05899*.

IXI 2015. 2015. IXI Dataset. https://brain-development.org/ixi-dataset/.

Jack Jr, C. R.; Bernstein, M. A.; Fox, N. C.; Thompson, P.; Alexander, G.; Harvey, D.; Borowski, B.; Britson, P. J.; L. Whitwell, J.; Ward, C.; et al. 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4): 685–691.

Ji, Y.; Bai, H.; Yang, J.; Ge, C.; Zhu, Y.; Zhang, R.; Li, Z.; Zhang, L.; Ma, W.; Wan, X.; et al. 2022. AMOS:

A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. *arXiv preprint arXiv:2206.08023*.

Kuijf, H. J.; Biesbroek, J. M.; De Bresser, J.; Heinen, R.; Andermatt, S.; Bento, M.; Berseth, M.; Belyaev, M.; Cardoso, M. J.; Casamitjana, A.; et al. 2019. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE transactions on medical imaging*, 38(11): 2556–2568.

Liew, S.-L.; Lo, B. P.; Donnelly, M. R.; Zavaliangos-Petropulu, A.; Jeong, J. N.; Barisano, G.; Hutton, A.; Simon, J. P.; Juliano, J. M.; Suri, A.; et al. 2022. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific data*, 9(1): 320.

Litjens, G.; Toth, R.; Van De Ven, W.; Hoeks, C.; Kerkstra, S.; Van Ginneken, B.; Vincent, G.; Guillard, G.; Birbeck, N.; Zhang, J.; et al. 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Medical image analysis*, 18(2): 359–373.

Luo, X.; Li, Z.; Zhang, S.; Liao, W.; and Wang, G. 2024. Rethinking Abdominal Organ Segmentation (RAOS) in the clinical scenario: A robustness evaluation benchmark with challenging cases.

Ma, J.; Zhang, Y.; Gu, S.; Ge, C.; Mae, S.; Young, A.; Zhu, C.; Yang, X.; Meng, K.; Huang, Z.; et al. 2024. Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the FLARE22 challenge. *The Lancet Digital Health*, 6(11): e815–e826.

Marcus, D. S.; Wang, T. H.; Parker, J.; Csernansky, J. G.; Morris, J. C.; and Buckner, R. L. 2007. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9): 1498–1507.

Marek, K.; Jennings, D.; Lasch, S.; Siderowf, A.; Tanner, C.; Simuni, T.; Coffey, C.; Kieburtz, K.; Flagg, E.; Chowdhury, S.; et al. 2011. The Parkinson progression marker initiative (PPMI). *Progress in neurobiology*, 95(4): 629–635.

Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. 2014. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34(10): 1993–2024.

Nugent, A. C.; Thomas, A. G.; Mahoney, M.; Gibbons, A.; Smith, J. T.; Charles, A. J.; Shaw, J. S.; Stout, J. D.; Namyst, A. M.; Basavaraj, A.; et al. 2022. The NIMH intramural healthy volunteer dataset: A comprehensive MEG, MRI, and behavioral resource. *Scientific Data*, 9(1): 518.

Reddy, G. 2023. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. *arXiv preprint arXiv:2312.03002*.

Rosenhain, S.; Magnuska, Z. A.; Yamoah, G. G.; Rawashdeh, W.; Kiessling, F.; Gremse, F.; et al. 2018. A preclinical micro-computed tomography database including 3D whole body organ segmentations. *Scientific data*, 5(1): 1–9.

Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; et al. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3): e1001779.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37: 84839–84865.

Wang, X.; Wang, W.; Cao, Y.; Shen, C.; and Huang, T. 2023a. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6830–6839.

Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023b. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*.

Wasserthal, J.; Breit, H.-C.; Meyer, M. T.; Pradella, M.; Hinck, D.; Sauter, A. W.; Heye, T.; Boll, D. T.; Cyriac, J.; Yang, S.; et al. 2023. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5): e230024.

Wu, J.; and Xu, M. 2024. One-prompt to segment all medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11302–11312.

Yang, K.; Musio, F.; Ma, Y.; Juchler, N.; Paetzold, J. C.; Al-Maskari, R.; Höher, L.; Li, H. B.; Hamamci, I. E.; Sekuboyina, A.; et al. 2023. Benchmarking the cow with the topcow challenge: Topology-aware anatomical segmentation of the circle of willis for cta and mra. *ArXiv*.

Yang, Y.; Su, G.; Hu, J.; Sammarco, F.; Geiping, J.; and Wolfers, T. 2025. Medsamix: A training-free model merging approach for medical image segmentation. *arXiv preprint arXiv:2508.11032*.

Yang, Y.; Ye, C.; Su, G.; Zhang, Z.; Chang, Z.; Chen, H.; Chan, P.; Yu, Y.; and Ma, T. 2024. BrainMass: Advancing Brain Network Analysis for Diagnosis with Large-scale Self-Supervised Learning. *IEEE Transactions on Medical Imaging*.

Zeng, A.; Wu, C.; Lin, G.; Xie, W.; Hong, J.; Huang, M.; Zhuang, J.; Bi, S.; Pan, D.; Ullah, N.; et al. 2023. ImageCAS: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images. *Computerized Medical Imaging and Graphics*, 109: 102287.

Zhang, Y.; Wang, J.; Pan, T.; Jiang, Q.; Ge, J.; Guo, X.; Jiang, C.; Lu, J.; Zhang, J.; Liu, X.; et al. 2024. Nasalseg: A dataset for automatic segmentation of nasal cavity and paranasal sinuses from 3d ct images. *Scientific Data*, 11(1): 1329.