Dual-Adaptive SAM3: Hierarchical Routing over Low-Rank Expert Layers for Parameter-Efficient Medical Image Segmentation

Ying Chen[1], Jinyue Li[2], Kun Wang[3†], Qiankun Li[4†], and Yang Liu[3]

[1] Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China
[2] University of Science and Technology of China
[3] Nanyang Technological University, Singapore
[4] Imperial Global Singapore (IGS), Imperial College London
kun.wang@ntu.edu.sg, q.li2@imperial.ac.uk
[†]Corresponding author.

Abstract. The Segment Anything Model with Concepts (SAM3) heralds a new paradigm for open-vocabulary segmentation through natural language interaction, offering significant potential for medical image analysis. However, effectively adapting such a powerful vision-language model to the diverse and nuanced domain of medical imaging remains a key challenge. Naive fine-tuning is parameter-inefficient, while standard Mixture-of-Experts (MoE) methods introduce prohibitive computational overhead, limiting their clinical applicability. To address this, we propose Dual-Adaptive SAM3 (DA-SAM3), a novel framework that achieves both high segmentation accuracy and extreme parameter efficiency via a dual-adaptive specialization mechanism. Our first adaptation is task-aware: a Dynamic Expert Router (DER) that sparsely activates the most relevant experts by jointly reasoning about the visual input and the textual concept prompt, mimicking a clinical consultation process. Our second adaptation is parameter-aware: a Decomposed Parameterized Experts (DPE) design that represents each expert as a shared frozen base (inherited from the pretrained SAM3) and a lightweight trainable low-rank delta, reducing MoE parameter overhead by over **80%**. Extensive experiments on multiple public medical segmentation benchmarks demonstrate that Dual-Adaptive SAM3 not only matches or exceeds the accuracy of fully fine-tuned SAM3 and standard MoE baselines, but also achieves a notable **5%** gain over current state-of-the-art methods, with interpretable results validating its effectiveness. The code is available at: https://github.com/Reconsider80/DA-SAM3.

Keywords: Segment Anything Model 3· Mixture-of-Experts· Parameter-Efficient Fine-Tuning· Medical Image Segmentation

1 Introduction

Medical image segmentation is fundamental to diagnostics and treatment planning, enabling critical applications from tumor delineation to organ analysis.

The rise of large-scale vision-language foundation models, such as the Segment Anything Model (SAM) [12] and its conceptual extension SAM3 [2], introduces a paradigm shift by supporting natural language-guided segmentation (e.g., "segment the left ventricle"). This interactivity promises more intuitive clinician-AI collaboration and alleviates reliance on large annotated medical datasets [10]. However, deploying such general-purpose models in the highly specialized domain of medical imaging presents a critical efficiency-specialization trade-off. The heterogeneity of imaging modalities, anatomical regions, and pathological appearances demands adaptable representations [20], yet common adaptation strategies are ill-suited: full fine-tuning is parameter-inefficient and risks catastrophic forgetting of valuable pretrained knowledge, while conventional Mixture-of-Experts (MoE) models introduce prohibitive parameter and memory overhead [22], limiting their scalability in resource-constrained clinical settings.

In pursuit of parameter efficiency, recent Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA [8] and Adapters offer a compelling alternative. However, their typical static and uniform application across the model fails to address the dynamic and hierarchical nature of medical concept understanding. Concurrently, novel efficient MoE designs like DeRS [9] demonstrate the potential of "upcycling" pretrained weights into conditional experts with minimal overhead. Despite these parallel advances, a crucial gap remains: how to integrate dynamic, input-conditioned routing with extreme parameter efficiency within a multimodal vision-language framework for medical segmentation. Thus, our core research question is: How can we efficiently specialize a large vision-language model like SAM3 to diverse medical concepts in a parameter-aware and task-aware manner, without sacrificing accuracy or practical deployability?

To bridge this gap, we propose Dual-Adaptive SAM3 (DA-SAM3), a novel framework that synergizes two complementary adaptive mechanisms for efficient medical image segmentation. First, a task-aware Dynamic Expert Router (DER) performs sparse, multimodal expert selection by jointly reasoning about visual content and textual concepts, emulating a context-sensitive clinical consultation. Second, a parameter-aware Decomposed Parameterized Experts (DPE) design decomposes each expert into a shared frozen based inherited from the pretrained SAM3 and lightweight low-rank "delta" matrices, inspired by efficient upcycling paradigms. Crucially, these dual mechanisms are orchestrated within a hierarchical specialization strategy, where experts at different network depths learn to specialize in coarse alignment, semantic identification, and boundary refinement, mirroring a coarse-to-fine clinical reasoning process. DA-SAM3 achieving an approximately **5%** improvement over state-of-the-art methods but also reduces MoE parameter overhead by over **80%**. Our contributions are threefold:

(1) We propose Dual-Adaptive SAM3, the first framework to integrate dynamic multimodal routing with parameter-efficient expert decomposition for adapting vision-language foundation models to medical imaging.

(2) We introduce the Dynamic Expert Router (DER) for context-aware sparse expert selection, and Decomposed Parameterized Experts (DPE), a novel low-rank parameterization that drastically reduces MoE parameter overhead.

(3) Through comprehensive experiments on multiple public benchmarks, we demonstrate that our approach matches or surpasses the accuracy of standard fine-tuning and MoE baselines, while significantly improving parameter and computational efficiency providing a practical pathway toward deploying versatile, generalist AI models in clinical practice.

2 Method

2.1 Overview

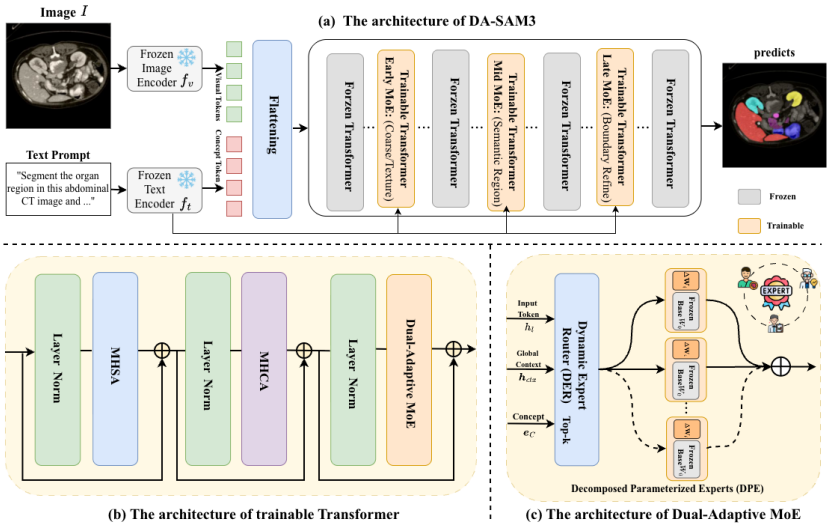

Fig. 1. **(a)** The architecture of the Dual-Adaptive SAM3 framework. **(b)** The architecture of trainable Transformer. **(c)** The Dual-Adaptive MoE Layer including Dynamic Expert Router **(DER)** and Decomposed Parameterized Expert **(DPE)**

.

As shown in Fig.1, we propose Dual-Adaptive SAM3 (DA-SAM3), a parameter-efficient framework for medical image segmentation. DA-SAM3 dynamically specializes the Segment Anything Model with Concepts (SAM3) via a hierarchical Mixture-of-Experts (MoE) adaptation. The core innovation is the targeted replacement of Feed-Forward Network (FFN) blocks within SAM3's fusion module with our Dual-Adaptive MoE Layers, enabling context-aware feature specialization guided by both visual characteristics and textual prompts. DA-SAM3 adapts the concept-conditioned SAM3 architecture through three main components: (1) a frozen image encoder f_v extracting visual features V; (2) a frozen text encoder f_t providing concept embeddings e_c; and (3) a trainable Transformer-based fusion module. To maintain parameter efficiency and preserve pre-trained general

knowledge, f_v and f_t remain entirely frozen. We exclusively fine-tune the newly inserted Dual-Adaptive MoE layers and Layer Normalization (LN) parameters within the fusion module. All other components, including the multi-head self-attention (MHSA) layers and Multi-Scale Cross-Attention (MSCA) layers, are kept frozen. This selective optimization strategy ensures dynamic domain adaptation while minimizing the trainable parameter footprint.

2.2 Hierarchical Fusion Decoder

The Hierarchical Fusion Decoder is a lightweight Transformer-based module designed to ground the linguistic [SEG] query into the visual domain. Unlike standard decoders that treat all layers uniformly, DA-SAM3 strategically substitute standard FFNs with **Dual-Adaptive MoE (DA-MoE)** layers at critical depths $\{L/6, L/4, L/2\}$. This hierarchical placement is not merely a structural choice but is intentionally designed to emulate the **multi-stage clinical reasoning process** typically employed by radiologists:

Early Stage (Coarse Global Alignment, $L/6$ **)**: Mirroring the "initial survey" in clinical reading, the MoE leverages global context $\mathbf{h}_{ctx}$ to align low-level visual primitives with the input concept.

Mid Stage (Semantic Structure Identification, $L/4$ **)**: At the intermediate depth, the model transitions to "pattern recognition." The experts focus on integrating multi-modal features to resolve complex morphological variations and identify organ-specific textures. This stage ensures the model correctly distinguishes between adjacent structures with similar intensities.

Late Stage (Precision Boundary Refinement, $L/2$ **)**: Equivalent to the "fine-grained contouring" required for surgical planning, these experts operate on high-resolution feature maps. The routing mechanism prioritizes experts capable of sharpening boundaries and resolving regional ambiguities, ensuring the final mask captures subtle pathological extensions or thin tissue interfaces.

The module processes the visual feature map $V \in \mathbb{R}^{H\times W\times D_v}$ and concept embedding $e_c \in \mathbb{R}^{D_t}$ into a unified sequence $\mathbf{X}_0$:

$$\mathbf{X}_0 = [\mathbf{V}\text{tok}; \mathbf{C}\text{tok}; \mathbf{S}\text{tok}] + \mathbf{P} \in \mathbb{R}^{(N_v+2)\times D}, \tag{1}$$

where $\mathbf{V}_{\text{tok}}$, $\mathbf{C}_{\text{tok}}$, and $\mathbf{S}_{\text{tok}}$ represent flattened visual features, projected concept embeddings, and a learnable segmentation query, respectively. $\mathbf{P}$ denotes learnable positional encodings. The sequence $\mathbf{X}_0$ is propagated through L layers.

2.3 Dual-Adaptive MoE Layer

We replace the standard FFN in selected layers of decoder with our proposed Dual-Adaptive MoE Layer. For a target layer l receiving input features $\mathbf{H}_l$, this layer operates as follows: **Dynamic Expert Router (DER)** The router generates sparse, token-wise gating weights by conditioning on a dual signal: a global domain-context vector and the local token features. Domain-Context Vector $\mathbf{h}_{ctx}$: To inform the router about the global image domain and its relevance to the

concept, we compute a compact summary. The concept token $\mathbf{C}_{tok}$ attends to a spatially pooled version of the visual features:

$$\mathbf{h}_{ctx} = \text{CrossAttn}\left(\mathbf{C}_{tok}, \text{Pool}(\mathbf{V})\right). \tag{2}$$

Routing Score Calculation: For the j-th token $\mathbf{h}_l^j$ in $\mathbf{H}_l$, the score for expert i is computed by fusing the token's content, the global context, and the original concept semantics:

$$s_i^j = \mathbf{W}_r \cdot \left[\mathbf{h}_l^j; \mathbf{h}_{ctx}; \text{StopGrad}\left(e_c\right)\right], \tag{3}$$

where $\mathbf{W}_r$ is a learnable projection and StopGrad ensures the router does not distort the frozen concept embedding. Sparse Top-k Gating: A sparse gating function selects the top-k experts per token for efficient computation, producing routing weights p_i^j. A sparse gating function selects the top- k experts (we use $k = 2$) per token for efficient computation, producing routing weights p_i^j. **Decomposed Parameterized Expert (DPE)** Inspired by DeRS, we parameterize each expert E_i not as a full dense network, but as a shared base weight plus a lightweight expert-specific delta. The original pre-trained FFN weights $\mathbf{W}_0$ from the corresponding SAM3 layer are retained as a shared, frozen base. Each expert i is defined by a lightweight, learnable low-rank delta:

$$\Delta\mathbf{W}_i = \mathbf{A}_i\mathbf{B}_i^\top, \quad \mathbf{A}_i \in \mathbb{R}^{D\times r}, \mathbf{B}_i \in \mathbb{R}^{D_{ff}\times r}, r \ll D. \tag{4}$$

The expert's operation is:

$$E_i(\mathbf{x}) = \left(\mathbf{W}_0 + \Delta\mathbf{W}_i\right)\mathbf{x}. \tag{5}$$

The output for token j is:

$$\mathbf{y}_l^j = \sum_{i\in\text{TopK}} p_i^j \cdot \text{GELU}\left(E_i\left(\mathbf{h}_l^j\right)\right). \tag{6}$$

This design introduces minimal new parameters while enabling dynamic specialization.

2.4 Two-Stage Specialization

We employ a two-stage training strategy to ensure stable expert specialization and effective routing. **Warm-up Stage (Expert Specialization)**: We initialize the low-rank delta matrices $\{\mathbf{A}_i, \mathbf{B}_i\}$ to zero and the router randomly. The model is first trained on a diverse medical segmentation dataset with a standard segmentation loss $\mathcal{L}_{\text{seg}}$. An auxiliary Load Balancing Loss $\mathcal{L}_{\text{balance}}$ is added to prevent router collapse and ensure equitable expert utilization across the batch. **Fine-tuning Stage (Routing Calibration)**: We then fine-tune the router parameters (with experts fixed) on a broader dataset that includes challenging or

composite concepts. This stage enhances the model's open-vocabulary generalization. The final training objective is:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{Focal Loss}}, \mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda_1 \mathcal{L}_{\text{balance}} + \lambda_2 \mathcal{L}_{\text{sparse}}, \tag{7}$$

where $\mathcal{L}_{\text{sparse}}$ is a regularization term encouraging the sparsity of the routing weights.

3 Experiments

3.1 Datasets

We evaluate DA-SAM3 on four widely adopted public datasets spanning cardiac and abdominal imaging. The **Synapse** dataset [5] comprises 30 abdominal CT scans, with 18 allocated to training and 12 to testing, featuring eight abdominal organ labels. The **MMWHS** dataset [25] contains 20 cardiac CT volumes, which we split into 16 training and 4 testing instances following the original protocol. The **BTCV** dataset [13] provides 30 annotated CT volumes covering 13 abdominal structures; we adopt its standard split of 24 training and 6 testing cases. The **ACDC** dataset [1] contributes 150 cardiac MRI acquisitions, all utilized with their predefined training-test partitions. For all experiments, we adhere to the official data splits to ensure fair comparison with prior work.

3.2 Implementation Details and Evaluation Metrics

To facilitate training, we apply various data augmentation techniques, including flipping, rotation, scaling, and intensity shifting. All images, except those from the Synapse CT dataset, are resized to 256×256, while Synapse CT images are resized to 224×224. The model is trained with a batch size of 8 using the AdamW optimizer, and weight decay = 0.1. The hyper-parameters λ_1 and λ_2 were empirically set to 0.01 and 0.001, respectively. This configuration follows the established practice in sparse MoE training, ensuring a balanced expert utilization and sharp routing decisions without compromising the convergence of the primary segmentation objective $\mathcal{L}_{seg}$. The learning rate is set to 0.0005, and a warmup strategy is employed to ensure stable convergence during the early stages. In the MoE configuration, we set the number of experts to 4 and the top-k value to half of the total feature count.

3.3 Comparison with State-of-the-Art Methods

To validate the efficacy of the proposed DA-SAM3, we conducted extensive benchmarks against contemporary SOTA SAM-based variants and task-specific architectures across four standardized datasets. These SAM-derived frameworks are bifurcated into prompt-guided and prompt-agnostic categories. For the former, a single point, stochastically sampled from the gold-standard mask, serves as the input prompt.Quantitative evidence in Table 1 reveals that DA-SAM3

establishes a new performance ceiling, markedly eclipsing existing SAM adaptations in both Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD). Furthermore, as detailed in the upper section of Table 1, our model consistently outperforms domain-specialized networks, underscoring its robust competitive edge in medical scenarios.

Table 1. Comparison with state-of-the-art methods on Synapse CT, MMWHS, BTCV, and ACDC.

Category	Method	Synapse CT		MMWHS		BTCV		ACDC	
		DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓
Task-specific	nnU-Net [11]	79.89	28.520	87.55	17.720	72.67	15.720	91.54	1.086
	TransUNet [3]	79.95	11.580	88.47	24.310	77.67	9.498	88.10	1.538
	Swin-UNETR [7]	80.58	15.460	88.92	14.310	78.11	7.405	89.74	1.239
	MedNeXt [18]	82.69	11.980	88.55	14.950	80.81	7.379	90.88	1.129
	Swin-UMamba [16]	83.48	8.140	88.91	15.060	80.59	5.910	90.39	1.253
Prompt-free SAM	SAMed [24]	80.42	10.770	87.050	25.320	71.23	9.010	88.83	1.429
	AutoSAM [19]	81.61	10.170	88.71	12.990	75.77	7.693	72.05	3.248
	H-SAM [4]	80.27	13.170	87.33	14.110	72.81	7.037	88.38	1.410
	MoE-SAM [14]	84.71	8.756	89.38	13.670	76.82	5.637	91.89	1.064
Prompt-based SAM	SAM [12]	64.94	39.830	82.11	46.940	63.84	20.360	75.15	4.311
	MedSAM [17]	72.45	20.430	84.53	55.940	69.14	18.490	82.11	3.720
	MSA [21]	77.13	25.340	85.50	35.680	72.31	17.51	83.01	2.715
	SAMUS [15]	70.55	43.650	83.98	30.740	65.12	24.58	69.66	5.559
	DeSAM [6]	76.77	9.704	81.54	16.340	68.08	7.263	67.26	5.391
	SAM-Med2D [23]	66.96	22.850	81.42	59.660	53.64	23.630	80.02	4.587
	SAM3 [2]	80.75	15.120	85.10	18.100	72.24	8.080	85.03	5.280
	DA-SAM3 (Ours)	**85.12**	**8.425**	**90.06**	**12.880**	**77.35**	**5.587**	**91.93**	**1.064**

As depicted in Fig. 2, our method exhibits precise alignment with actual organ structures, thereby validating the efficacy of our location-aware design. In contrast, both MoE-SAM and SAM3 frequently misclassify different organs, underscoring the superiority of our Dual-Adaptive MoE Layer strategy.

3.4 Ablation Study

Table 2. Ablation study of module's key component on three datasets.

Category	Method	Synapse CT		MMWHS		ACDC	
		DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓
Baseline	SAM3	80.75	15.120	85.10	18.100	72.24	8.080
Fusion Strategy Variants	SAM3+Concat	80.82	15.090	88.18	18.130	73.23	8.106
Fine-tuning Strategy Variants	SAM3+LoRA	82.82	14.342	83.28	15.482	73.35	7.231
	SAM3+Standard MoE	82.28	14.090	85.18	16.630	73.03	7.480
	Ours w/o DER (Random Route)	82.75	12.120	86.10	14.100	73.24	7.080
	Ours w/o DPE (Full Expert)	83.55	11.700	86.60	13.700	74.55	6.102
Ours	DA-SAM3 (Ours Full)	**85.12**	**8.425**	**90.06**	**12.880**	**77.35**	**5.587**

Table 2 presents ablation results on three datasets. Compared to the standard SAM3 baseline, our full model achieves substantial gains, notably improving DSC by 4.37% on Synapse CT. While SAM3+LoRA and Standard MoE offer

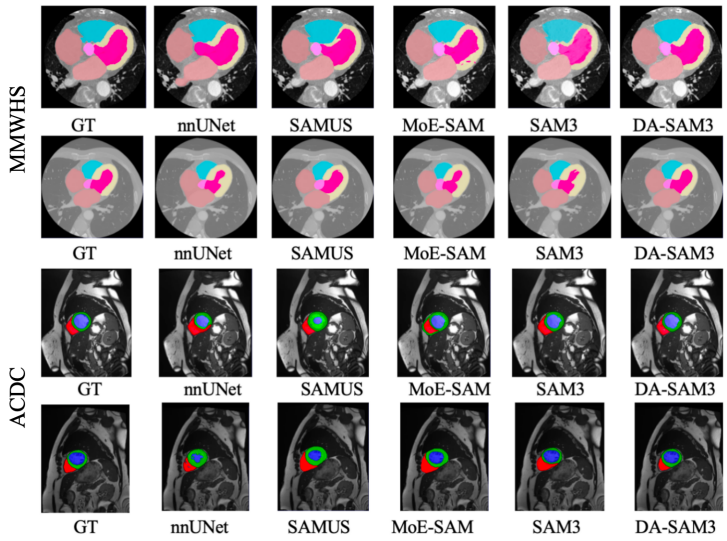

Fig. 2. Visual comparison of segmentation results.

incremental improvements, they fall short of DA-SAM3's performance. Specifically, our method reduces the HD score from 15.120mm to 8.425mm on Synapse, demonstrating that targeted expert specialization at the semantic bottleneck is more effective than global low-rank adaptations or unconstrained expert layers for capturing precise anatomical boundaries. The SAM3+Concat (where visual and text tokens are simply concatenated without hierarchical interaction) variant shows limited efficacy, confirming that simple feature aggregation cannot resolve complex cross-modal misalignments. Furthermore, replacing our DER with Random Routing leads to a significant performance drop. This underscores the router's critical role in utilizing clinical concept embeddings to dynamically activate task-relevant experts. Ablating the DPE results in inferior performance across all metrics. This validates that our parametric decomposition acts as a robust regularizer, preventing expert over-parameterization while maximizing the model's capacity to resolve ambiguous clinical textures.

4 Conclusion

In this paper, we introduce Dual-Adaptive SAM3 (DA-SAM3), a novel framework that achieves parameter-efficient medical image segmentation through hierarchical routing over low-rank expert layers. DA-SAM3 fundamentally rethinks adaptation by expanding a single, static adapter into multiple, dynamically selectable expert adapters, enabling the model to perform input-conditioned, coarse-to-fine reasoning that mimics clinical decision-making. A Dynamic Expert Router (DER) that sparsely activates the most relevant experts by jointly

reasoning about visual content and textual concepts, and Decomposed Parameterized Experts (DPE) that decompose each expert into a shared frozen base and lightweight low-rank deltas. Extensive experiments on four public benchmark datasets demonstrate that DA-SAM3 not only matches but often surpasses the accuracy of fully fine-tuned models and standard MoE baselines, while maintaining extreme parameter efficiency.

Acknowledgments. This research is part of the IN-CYPHER programme and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

Disclosure of Interests. The authors have no competing interests in the paper.

References

1. Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.
2. Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.
3. Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
4. Zhiheng Cheng, Qingyue Wei, Hongru Zhu, Yan Wang, Liangqiong Qu, Wei Shao, and Yuyin Zhou. Unleashing the potential of sam for medical adaptation via hierarchical decoding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3511–3522, 2024.
5. Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020.
6. Yifan Gao, Wei Xia, Dingdu Hu, Wenkui Wang, and Xin Gao. Desam: Decoupled segment anything model for generalizable medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 509–519. Springer, 2024.
7. Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.
8. Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
9. Yongqi Huang, Peng Ye, Chenyu Huang, Jianjian Cao, Lin Zhang, Baopu Li, Gang Yu, and Tao Chen. Ders: Towards extremely efficient upcycled mixture-of-experts models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10056–10066, 2025.

10. Ziyan Huang, Zhongying Deng, Jin Ye, Haoyu Wang, Yanzhou Su, Tianbin Li, Hui Sun, Junlong Cheng, Jianpin Chen, Junjun He, et al. A-eval: A benchmark for cross-dataset and cross-modality evaluation of abdominal multi-organ segmentation. *Medical Image Analysis*, 101:103499, 2025.
11. Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
12. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.
13. Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, page 12. Munich, Germany, 2015.
14. Ruocheng Li, Lei Wu, Jingjun Gu, Qi Xu, Wanyi Chen, Xiaoxu Cai, and Jiajun Bu. Moe-sam: Enhancing sam for medical image segmentation with mixture-of-experts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 367–377. Springer, 2025.
15. Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–34. Springer, 2024.
16. Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International conference on medical image computing and computer-assisted intervention*, pages 615–625. Springer, 2024.
17. Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.
18. Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 405–415. Springer, 2023.
19. Tal Shaharabany, Aviad Dahan, Raja Giryes, and Lior Wolf. Autosam: Adapting sam to medical images by overloading the prompt encoder. *arXiv preprint arXiv:2306.06370*, 2023.
20. Jianghao Wu, Xinya Liu, Guotai Wang, and Shaoting Zhang. Sictta: Single image continual test time adaptation for medical image segmentation. *Medical Image Analysis*, 108:103859, 2026.
21. Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547, 2025.
22. Xian-Tao Wu, Xiao-Diao Chen, Wen Wu, Weiyin Ma, and Haichuan Song. Accelerating vision foundation model for efficient medical image segmentation. *Medical Physics*, 53(1):e70193, 2026.
23. Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*, 2023.

24. Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.
25. Xiahai Zhuang and Juan Shen. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Medical image analysis*, 31:77–87, 2016.