SafeClick: Error-Tolerant Interactive Segmentation of Any Medical Volumes via Hierarchical Expert Consensus

Yifan Gao[1,2,3*], Jiaxi Sheng[1,3*], Wenbin Wu[1,3], Haoyue Li[1,4], Yaoxian Dong[1,3], Chaoyang Ge[1,3], Feng Yuan[1,3], Xin Gao[2,3✉]

[1] School of Biomedical Engineering (Suzhou), Division of Life Science and Medicine, University of Science and Technology of China, Hefei, China
[2] Shanghai Innovation Institute, Shanghai, China
[3] Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, China
[4] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China

Abstract. Foundation models for volumetric medical image segmentation have emerged as powerful tools in clinical workflows, enabling radiologists to delineate regions of interest through intuitive clicks. While these models demonstrate promising capabilities in segmenting previously unseen anatomical structures, their performance is strongly influenced by prompt quality. In clinical settings, radiologists often provide suboptimal prompts, which affects segmentation reliability and accuracy. To address this limitation, we present SafeClick, an error-tolerant interactive segmentation approach for medical volumes based on hierarchical expert consensus. SafeClick operates as a plug-and-play module compatible with foundation models including SAM 2 and MedSAM 2. The framework consists of two key components: a collaborative expert layer (CEL) that generates diverse feature representations through specialized transformer modules, and a consensus reasoning layer (CRL) that performs cross-referencing and adaptive integration of these features. This architecture transforms the segmentation process from a prompt-dependent operation to a robust framework capable of producing accurate results despite imperfect user inputs. Extensive experiments across 15 public datasets demonstrate that our plug-and-play approach consistently improves the performance of base foundation models, with particularly significant gains when working with imperfect prompts. The source code is available at https://github.com/yifangao112/SafeClick.

Keywords: Interactive Medical Image Segmentation · Foundation Model · Segment Anything Model 2 · Error-Tolerant

[*] These authors contributed equally to this work.
[✉] Corresponding author

1 Introduction

Medical image segmentation plays a vital role in clinical applications, facilitating precise delineation of regions of interest within medical images [1–3]. Recent advancements in vision foundation models have enabled radiologists to segment these regions through intuitive prompts, such as points and bounding boxes. These interactive approaches significantly streamline diagnostic and therapeutic workflows [4, 5].

The evolution of foundation models for volumetric medical image segmentation has expanded the capabilities of automated analysis beyond 2D images. Models like SAM 2 [6] and MedSAM 2 [7] have demonstrated promising performance across various medical imaging tasks [8–11]. However, a significant limitation emerges when these models receive imperfect prompts, leading to substantial performance degradation [12]. Studies have shown that even minor deviations in prompt placement can result in up to a 30% decrease in segmentation performance [13]. Similarly, research indicates performance drops by nearly 20% when point prompts are randomly selected instead of precisely placed [14]. This sensitivity to prompt quality severely restricts the clinical applicability of such models, as radiologists often struggle to provide consistently perfect prompts in the fast-paced and complex clinical environment.

To address this challenge, we introduce SafeClick, an error-tolerant interactive segmentation approach for medical volumes based on hierarchical expert consensus. SafeClick transitions from reliance on a single prompt-driven process to a collaborative decision-making framework. Our approach consists of two key components: the collaborative expert layer (CEL) and the consensus reasoning layer (CRL). The CEL comprises three specialized transformer modules: one processing intermediate image features, another analyzing final image features independently of prompts, and a third integrating prompt information. The CRL then dynamically fuses these complementary perspectives through cross-referencing and feature aggregation, emphasizing the most reliable features when prompt quality varies. This hierarchical consensus mechanism enables SafeClick to maintain high performance across diverse prompt conditions, enhancing reliability in clinical settings.

Extensive experiments across 15 public datasets with diverse anatomical structures demonstrate that SafeClick outperforms state-of-the-art foundation models, achieving superior performance with both ideal and imperfect prompts. As visualized in Fig. 1, SafeClick consistently outperforms baseline foundation models across multiple datasets, demonstrating its robust generalization capabilities across diverse anatomical structures. Our contributions are summarized as follows: (1) We present a plug-and-play module that enhances foundation models' resilience against imperfect prompts without requiring architectural modifications; (2) We introduce a hierarchical expert consensus mechanism that effectively balances prompt-dependent and image-intrinsic features; (3) We demonstrate consistent performance improvements across diverse anatomical structures and imaging modalities, establishing SafeClick's broad applicability in clinical scenarios.

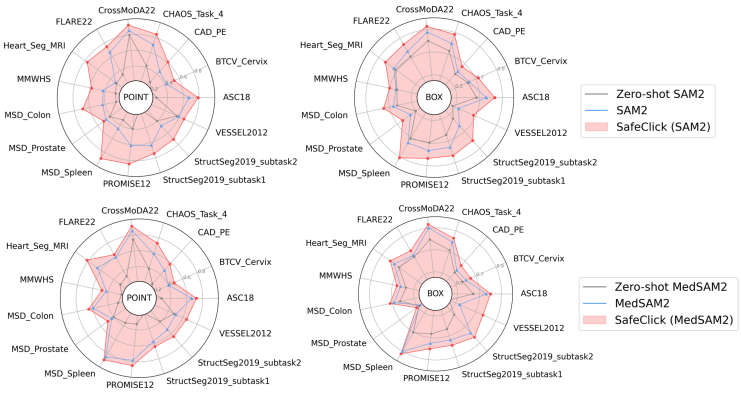

Fig. 1. Performance comparison between SafeClick and baseline foundation models across different datasets.

2 Methodology

Fig. 2 presents the overall architecture of our proposed module. It adopts the same encoder as recent foundation models for medical volume segmentation, leveraging pre-trained weights for efficient feature extraction. The decoder of SafeClick is distinctively composed of two complementary components: a collaborative expert layer (CEL) and a consensus reasoning layer (CRL). The CEL is tasked with analyzing and refining features across different representational spaces, comprising three expert layers equipped with self-attention or cross-attention mechanisms. In parallel, the CRL integrates multi-level features through cross-referencing and feature aggregation, optimizing the final segmentation output.

Collaborative Expert Layer: The collaborative expert paradigm draws from the principle of divide-and-conquer [15], breaking down complex segmentation tasks into manageable components that are addressed by specialized experts. Rather than relying solely on prompt-driven features, SafeClick distributes responsibility across multiple expert modules, each contributing complementary information to the segmentation process.

As shown in Fig. 2, the designed CEL within our framework is composed of three distinct transformer layers, each serving a specialized role in the process of feature analysis and refinement. This network includes two cross-attention transformer layers, designated as E_1 and E_3, alongside a self-attention transformer layer, referred to as E_2. Notably, the architectural design of E_3 mirrors the transformer layer found within the original mask decoder of foundation models.

Given inputs from the image encoder ($x_i \in \mathbb{R}^{H_p \times W_p \times C}$ and $x_f \in \mathbb{R}^{H \times W \times C}$) and prompt encoder ($x_p \in \mathbb{R}^{H \times W \times C}$), the CEL processes these representations through its specialized components. We select the $m/2$-th layer output from the m-layer image encoder as x_i to capture multi-scale information.

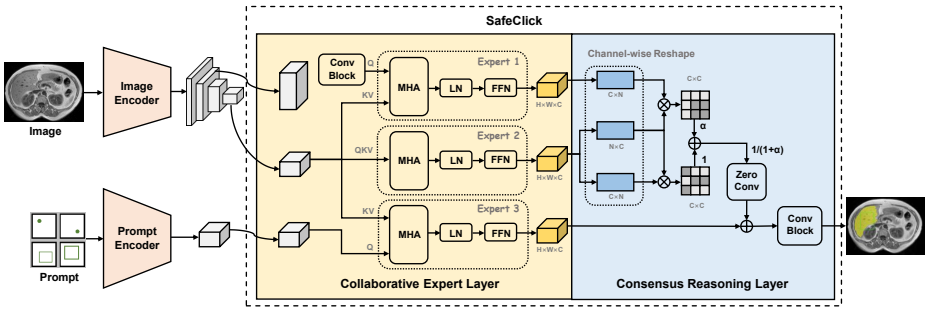

Fig. 2. Architecture of the proposed SafeClick. The diagram illustrates the two primary components: the collaborative expert layer (CEL) that generates diverse feature representations through specialized transformer modules, and the consensus reasoning layer (CRL) that performs cross-referencing and adaptive integration of these features. The framework operates as a plug-and-play module compatible with foundation models for medical volume segmentation.

The intermediate feature x_i first undergoes a dimension transformation to match the spatial resolution of other features:

$$\hat{x}_i = \mathcal{F}_{transform}(x_i) \in \mathbb{R}^{H \times W \times C} \tag{1}$$

where $\mathcal{F}_{transform}$ is a sequential operation including convolution, normalization, and activation. The first expert E_1 applies cross-attention between transformed intermediate and final image features:

$$\tilde{x}_1 = \mathrm{MCA}\big(\mathrm{LN}\big(\mathcal{T}_{cl}(\hat{x}_i)\big), \mathrm{LN}\big(\mathcal{T}_{cl}(x_f)\big)\big) + \mathcal{T}_{cl}(\hat{x}_i) \tag{2}$$

$$x_1 = \mathcal{R}\big(\mathrm{MLP}(\mathrm{LN}(\tilde{x}_1)) + \tilde{x}_1\big) \tag{3}$$

where LN is layer normalization and MCA is multi-head cross-attention. $\mathcal{T}_{cl}(\cdot)$ is the channel last transformation, which transforms the spatial features with dimensions $H \times W \times C$ into a 2D embedding of $N \times C$ to capture spatial relationships in attention computation. $\mathcal{R}(\cdot)$ transform the 2D embedding back to the image dimensions. The second expert E_2 applies self-attention to final image features:

$$\tilde{x}_2 = \mathrm{MSA}\big(\mathrm{LN}\big(\mathcal{T}_{cl}(x_f)\big)\big) + \mathcal{T}_{cl}(x_f) \tag{4}$$

$$x_2 = \mathcal{R}\big(\mathrm{MLP}(\mathrm{LN}(\tilde{x}_2)) + \tilde{x}_2\big) \tag{5}$$

where MSA is multi-head self-attention, enabling prompt-independent analysis of image content.The third expert E_3 processes prompt features x_p with x_f using cross-attention, producing output x_3. Through this architectural composition, our CEL not only enriches the feature space available for segmentation tasks but also introduces a level of adaptability and specificity that is essential for handling the variability inherent in medical volumes and user prompts.

Consensus Reasoning Layer: The Consensus Reasoning Layer (CRL) distills complementary information from the expert modules to handle imperfect user prompts in medical volume segmentation. Given outputs $x_1, x_2 \in \mathbb{R}^{H\times W\times C_1}$ from our expert layers and prompt-dependent features $x_3 \in \mathbb{R}^{H\times W\times C}$, the CRL performs cross-referencing between these representations.

We first transform the feature tensors into channel-first representation:

$$\Phi_1 = \mathcal{T}(x_1) \in \mathbb{R}^{C\times N}, \quad \Phi_2 = \mathcal{T}(x_2) \in \mathbb{R}^{C\times N} \tag{6}$$

where $\mathcal{T}(\cdot)$ denotes reshaping operation and $N = H \times W$.

We compute and normalize cross-reference attention matrices through a contrastive mechanism:

$$\hat{\mathcal{A}}_{cross} = \max(\Phi_1\Phi_2^\top)\cdot \mathbb{1} - \Phi_1\Phi_2^\top, \quad \hat{\mathcal{A}}_{self} = \max(\Phi_2\Phi_2^\top)\cdot \mathbb{1} - \Phi_2\Phi_2^\top \tag{7}$$

where $\mathbb{1}(\cdot)$ denotes the all-one matrix.

These attention matrices are combined through a learnable parameter α and applied to modulate features:

$$x_2' = \mathcal{R}\left(\frac{\sigma(\hat{\mathcal{A}}_{self}) + \alpha\cdot\sigma(\hat{\mathcal{A}}_{cross})}{1+\alpha}\Phi_2 + \Phi_2\right) \tag{8}$$

where $\sigma(\cdot)$ is softmax operation and $\mathcal{R}(\cdot)$ reshapes back to spatial dimensions.

The prompt-guided features are integrated with this refined representation:

$$x_{enhanced} = \phi(\mathcal{N}(\mathcal{F}_{conv}(x_2') + x_3)) \tag{9}$$

where $\mathcal{F}_{conv}$ is a convolutional operation with zero-initialized parameters, ensuring the network initially preserves the prompt-driven features while gradually learning optimal feature integration during training. $\mathcal{N}(\cdot)$ is layer normalization, and $\phi(\cdot)$ represents leaky ReLU activation.

This consensus building process enables our framework to balance multiple expert perspectives, maintaining high segmentation performance even with suboptimal user inputs. The output $x_{enhanced}$ is passed to upsampling layers for generating the final segmentation mask.

3 Experiments and Results

Dataset and Preprocessing: We evaluate SafeClick on 15 public 3D medical datasets spanning various anatomical regions including brain, abdomen, lung, cardiac, urology, gynecology, and vasculature structures, as shown in Table 1. These datasets collectively contain over 89,000 annotated regions of interest across different imaging modalities. For preprocessing, we normalize intensity values to the range [0, 1] and resample all volumes to isotropic spacing [16]. We simulate imperfect prompts by adding noise to perfect prompts: for point prompts, we apply random displacements from the center of mass ranging from

Table 1. Overview of the 15 public datasets used for evaluation, organized by body region. The table shows the anatomical focus and number of regions of interest (ROIs) in each dataset, covering a diverse range of medical imaging applications.

Dataset	Body	ROIs	Dataset	Body	ROIs	Dataset	Body	ROIs
ASC18 [17]	Brain	8855	FLARE22 [18]	Brain	23368	MSD_Spleen [19]	Abdomen	1008
BTCV_Cervix [20]	Gynecology	4667	Heart_Seg_MRI [21]	Cardiac	517	PROMISE12 [22]	Urology	776
CAD_PE [23]	Lung	3102	MMWHS [24]	Brain	17684	StructSeg2019_subtask1 [25]	Abdomen	3688
CHAOS_Task_4 [26]	Abdomen	3513	MSD_Colon [19]	Abdomen	1093	StructSeg2019_subtask2 [25]	Brain	4781
CrossMoDA22 [27]	Brain	1478	MSD_Prostate [19]	Urology	1466	VESSEL2012 [28]	Vasculature	13182

25% to 100% of the object radius; for bounding boxes, we scale the perfect box by factors ranging from 50% to 150%.

Implementation and Experiment Setting: We implement SafeClick as a plug-and-play module compatible with SAM 2 and MedSAM 2 across all their model sizes (ViT-T, ViT-S, ViT-L). For each dataset, we implemented random partitioning into training, validation, and test sets in a 7:1:2 ratio. Our implementation uses PyTorch, with training performed on NVIDIA H100 GPUs. We train with AdamW optimizer, using a learning rate of 1e-4 with cosine annealing schedule, batch size of 8, and train for 20 epochs. For a fair comparison, following the same approach as fine-tuning SAM2 and MedSAM2, we froze the encoder weights of the base model and trained only the SafeClick module. As a lightweight module, SafeClick introduces only an 18% increase in inference time compared to the baseline architectures. We evaluate performance using Dice similarity coefficient and compare results under both perfect and imperfect prompt conditions.

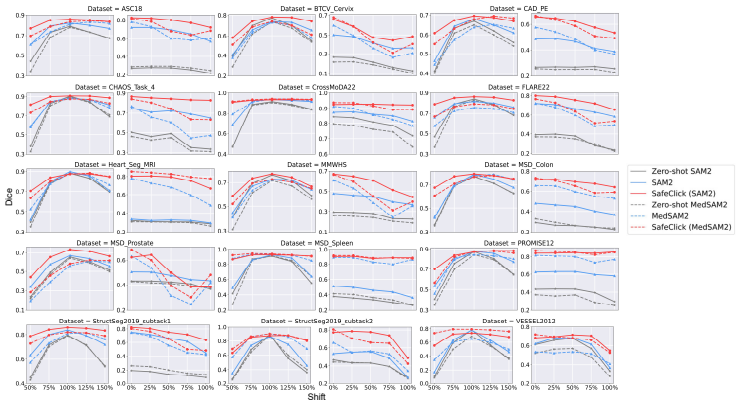

Fig. 3. Detailed results per dataset with box prompt (left) and point prompt (right).

Results: Table 2 and Fig. 3 report the quantitative comparison of various methods across 15 public datasets, with Fig. 4 offering visual comparisons of segmentation results. SafeClick consistently improves the performance of both

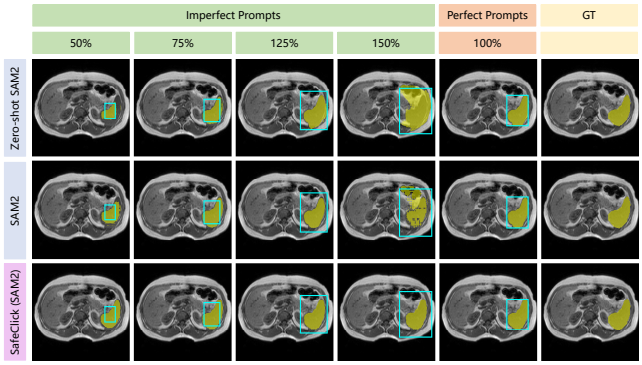

Fig. 4. Qualitative results in the CHAOS dataset under box prompts. The red star indicates the location of the point prompt. GT: ground truth.

Table 2. Performance comparison (Dice scores (%)) of SafeClick against baseline models under perfect prompts (PP) and imperfect prompts (IP) of varying quality. These results represent the average performance across 15 datasets. For point prompts, percentages indicate displacement from ideal position; for bounding boxes, percentages indicate scale relative to perfect box. Bold and underlined values indicate best and second-best results.

		Point						Bbox					
Methods	Model Types	PP	IP					PP	IP				
		0%	25%	50%	75%	100%	Avg.	100%	50%	75%	125%	150%	Avg.
Zero-shot SAM 2	ViT-T	44.08	43.72	42.69	38.57	33.79	39.69	81.33	35.15	71.90	73.20	60.68	60.23
	ViT-S	40.05	39.35	38.05	32.47	25.63	33.88	81.66	36.72	71.76	72.73	57.42	59.66
	ViT-L	35.15	34.17	33.75	30.58	25.65	31.04	81.25	40.51	73.16	72.16	59.59	61.36
SAM 2	ViT-T	56.06	55.85	54.58	51.10	42.76	51.07	81.53	40.52	72.59	76.08	65.66	63.71
	ViT-S	60.14	59.38	56.95	52.59	45.30	53.56	81.00	50.73	75.94	77.33	67.29	67.82
	ViT-L	61.68	61.00	58.86	54.44	49.27	55.89	82.36	53.52	76.02	79.72	70.21	69.87
SafeClick (SAM 2)	ViT-T	75.00	74.29	71.11	67.72	62.95	69.02	82.28	66.13	79.08	81.74	78.75	76.43
	ViT-S	78.06	77.33	74.03	70.88	**65.79**	72.01	83.57	**71.22**	**81.76**	82.93	**80.57**	**79.12**
	ViT-L	**78.38**	**77.94**	**74.66**	**71.09**	65.31	**72.25**	**83.88**	71.06	81.63	**82.95**	79.34	78.75
Improve		17.85	17.78	16.47	17.19	18.91	17.59	1.61	21.21	5.97	4.83	11.83	10.96
Zero-shot MedSAM 2	ViT-T	43.84	42.91	41.50	36.60	30.93	37.99	80.19	34.11	69.72	73.29	60.73	59.46
	ViT-S	40.29	39.69	38.01	33.99	26.75	34.61	78.84	31.06	67.10	70.88	57.66	56.68
	ViT-L	30.37	29.91	29.43	25.82	22.53	26.92	79.34	31.60	69.23	72.15	61.05	58.51
MedSAM 2	ViT-T	70.87	65.81	56.65	49.57	49.61	55.41	78.64	52.96	72.40	77.53	71.55	68.61
	ViT-S	72.23	67.72	61.42	53.99	53.14	59.07	79.94	53.70	73.53	79.18	74.00	70.10
	ViT-L	70.74	65.04	57.55	49.83	49.79	55.55	78.54	48.84	71.41	78.03	70.76	67.26
SafeClick (MedSAM 2)	ViT-T	77.75	73.52	65.34	57.54	58.52	63.73	80.62	64.48	76.67	80.21	77.04	74.60
	ViT-S	**79.06**	**75.01**	**68.22**	61.24	**61.64**	**66.53**	**81.29**	65.02	76.98	**81.09**	**78.55**	**75.41**
	ViT-L	78.41	74.79	67.23	**61.36**	60.91	66.07	81.06	**65.67**	**77.06**	80.78	77.32	75.21
Improve		7.13	8.25	8.39	8.92	9.51	8.77	1.95	13.22	4.46	2.45	5.53	6.42

SAM 2 and MedSAM 2 across all prompt conditions. For SAM 2, our method achieves average improvements of 17.59% and 10.96% in Dice score for imperfect point and bounding box prompts, respectively. For MedSAM 2, the average improvements for imperfect prompts are 8.77% and 6.42%. Notably, performance gains are most significant under challenging conditions with highly imperfect prompts, where baseline models experience substantial degradation.

For instance, with 50% displaced box prompts, SafeClick improves SAM 2's performance by 21.21%. The results demonstrate that our method effectively maintains segmentation quality even when prompt quality deteriorates, confirming its value in practical clinical scenarios where perfect user interaction cannot be guaranteed.

Ablation Study: We conduct ablation studies on the ASC18 dataset to evaluate the contribution of each component in our SafeClick framework, with results presented in Table 3. Removing the cross-attention transformer layer (E_1) results in performance drops of 2.93% and 1.44% for point and bounding box prompts under imperfect conditions, indicating its importance in cross-referencing multi-level features. When the self-attention transformer layer (E_2 is removed, performance decreases by 5.15% and 2.98%, suggesting this component's significant role in providing prompt-independent analysis of image content. Specially, the hybrid-attention transformer layer (E_3) refers to the Transformer layer of the original mask decoder in SAM2 to integrate the prompt with image features. Removing E_3 would prevent the model from processing the prompt and thus change the model's prompt interaction mode. As this paper focuses on researching prompts of different qualities, the E_3 expert layer is retained. The absence of the CRL leads to reductions of 4.53% and 2.30%, confirming its effectiveness in adaptively integrating expert outputs. These results demonstrate that each component contributes substantially to SafeClick's overall performance, with their collaborative interaction yielding optimal results.

Table 3. Ablation studies of the various components of Safeclick on ASC18 dataset.

Type		Baseline	w/o E_1	w/o E_2	w/o CRL	SafeClick		Baseline	w/o E_1	w/o E_2	w/o CRL	SafeClick
PP	Point	77.19	80.54	78.31	<u>80.99</u>	**83.46**	Bbox	83.05	84.20	83.46	<u>85.09</u>	**86.46**
IP Avg.		72.06	<u>76.98</u>	74.76	75.38	**79.91**		77.89	<u>81.55</u>	80.01	80.69	**82.99**

4 Discussion and Conclusion

We presented SafeClick, a plug-and-play error-tolerant module for interactive medical volume segmentation that significantly enhances foundation models' resilience against imperfect prompts. SafeClick's hierarchical expert consensus mechanism can be viewed from an ensemble learning perspective, where multiple specialized experts analyze the input data through complementary pathways, reducing the variance of predictions when faced with noisy inputs. This aligns with classical mixture-of-experts model where combining diverse estimators leads to improved performance.

Our extensive evaluation across 15 datasets shows interesting patterns regarding different prompt types. Point prompts, which provide minimal spatial guidance, benefit from consistent improvements regardless of prompt quality, as SafeClick's collaborative experts compensate for the limited information by

leveraging image features. Bounding box prompts, while generally more robust in baseline models, show particularly substantial gains under imperfect conditions (up to 21.21% improvement for severely distorted boxes), which is especially valuable in clinical settings where precise box placement is challenging during time-constrained examinations.

A notable strength of SafeClick is its universal compatibility with foundation models built upon SAM2 architecture without requiring major architectural modifications, enabling straightforward integration into existing clinical workflows across diverse imaging modalities. The module adds minimal computational overhead (approximately 18% additional time) while delivering substantial performance gains across all tested anatomical regions from brain to vasculature structures. While SafeClick significantly improves robustness, extremely poor prompts that completely miss the target region remain challenging. Future work could explore integrating SafeClick with automatic prompt correction mechanisms and extending its application to more diverse medical imaging types.

Acknowledgments. This work was supported by National Science Foundation of China under Grant 82372052.

Disclosure of Interests. The authors have no competing interests.

References

1. Yifan Gao, Yin Dai, Fayu Liu, Weibing Chen, and Lifu Shi. An anatomy-aware framework for automatic segmentation of parotid tumor from multimodal mri. *Computers in Biology and Medicine*, 161:107000, 2023.
2. Yifan Gao, Wei Xia, Wenkui Wang, and Xin Gao. Mba-net: Sam-driven bidirectional aggregation network for ovarian tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 437–447. Springer, 2024.
3. Yifan Gao, Yaoxian Dong, Wenbin Wu, Chaoyang Ge, Feng Yuan, Jiaxi Sheng, Haoyue Li, and Xin Gao. Wega: Weakly-supervised global-local affinity learning framework for lymph node metastasis prediction in rectal cancer. *arXiv preprint arXiv:2505.10502*, 2025.
4. Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7):1562–1573, 2018.
5. Xiangde Luo, Guotai Wang, Tao Song, Jingyang Zhang, Michael Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Medical image analysis*, 72:102102, 2021.
6. Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
7. Jiayuan Zhu, Yunli Qi, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024.

8. Tianrun Chen, Ankang Lu, Lanyun Zhu, Chaotao Ding, Chunan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2408.04579*, 2024.
9. Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *arXiv preprint arXiv:2408.08870*, 2024.
10. Yichi Zhang and Zhenrong Shen. Unleashing the potential of sam2 for biomedical images and videos: A survey. *arXiv preprint arXiv:2408.12889*, 2024.
11. Zhi Li, Kai Zhao, Yaqi Wang, and Shuai Wang. Adaptive interactive segmentation for multimodal medical imaging via selection engine. *arXiv preprint arXiv:2411.19447*, 2024.
12. 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.
13. Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024.
14. Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. Scribbleprompt: Fast and flexible interactive segmentation for any medical image. *arXiv preprint arXiv:2312.07381*, 2023.
15. Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.
16. 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.
17. Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67:101832, 2021.
18. Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023.
19. Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.
20. 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.
21. Catalina Tobon-Gomez, Arjan J Geers, Jochen Peters, Jürgen Weese, Karen Pinto, Rashed Karim, Mohammed Ammar, Abdelaziz Daoudi, Jan Margeta, Zulma Sandoval, et al. Benchmark for algorithms segmenting the left atrium from 3d ct and mri datasets. *IEEE transactions on medical imaging*, 34(7):1460–1473, 2015.

22. Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.
23. Germán González, Daniel Jimenez-Carretero, Sara Rodríguez-López, Carlos Cano-Espinosa, Miguel Cazorla, Tanya Agarwal, Vinit Agarwal, Nima Tajbakhsh, Michael B Gotway, Jianming Liang, et al. Computer aided detection for pulmonary embolism challenge (cad-pe). *arXiv preprint arXiv:2003.13440*, 2020.
24. Fuping Wu and Xiahai Zhuang. Minimizing estimated risks on unlabeled data: A new formulation for semi-supervised medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6021–6036, 2022.
25. Jun Shi. Structseg2019 gtv segmentation, 2023.
26. A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, and Sinem Aslan. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, April 2021.
27. Reuben Dorent, Aaron Kujawa, Marina Ivory, Spyridon Bakas, Nicola Rieke, Samuel Joutard, Ben Glocker, Jorge Cardoso, Marc Modat, Kayhan Batmanghelich, et al. Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. *Medical Image Analysis*, 83:102628, 2023.
28. Rina D Rudyanto, Sjoerd Kerkstra, Eva M Van Rikxoort, Catalin Fetita, Pierre-Yves Brillet, Christophe Lefevre, Wenzhe Xue, Xiangjun Zhu, Jianming Liang, Ilkay Öksüz, et al. Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the vessel12 study. *Medical image analysis*, 18(7):1217–1232, 2014.