TGSAM-2: Text-Guided Medical Image Segmentation using Segment Anything Model 2

Runtian Yuan[1], Ling Zhou[1], Jilan Xu[1], Qingqiu Li[1], Mohan Chen[1], Yuejie Zhang[1(✉)], Rui Feng[1(✉)], Tao Zhang[2], and Shang Gao[3]

[1] College of Computer Science and Artificial Intelligence, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
{yjzhang,fengrui}@fudan.edu.cn
[2] School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai, China
[3] School of Information Technology, Deakin University, Victoria, Australia

Abstract. The Segment Anything Model 2 (SAM-2) has shown impressive capabilities for promptable segmentation in images and videos. However, SAM-2 primarily operates on visual prompts including points, boxes, and masks, which does not natively support text prompts. This limitation is particularly noticeable in medical imaging, where domain-specific textual descriptions are often beneficial for annotating subtle abnormalities and identifying regions of interest. In this paper, we introduce Text-Guided SAM-2 (TGSAM-2), a medical image segmentation model tailored to leverage text prompts as contextual guidance. We propose a text-conditioned visual perception module that conditions visual features on textual descriptions, and refine the memory encoder to track target objects using medical text prompts. We evaluate our method on four medical image datasets with video-like characteristics, including 2D image sequences (e.g. Endoscopy, Ultrasound) and 3D volumes (e.g. CT, MRI). Experimental results demonstrate that our method outperforms state-of-the-art models, including both image-only and text-guided medical image segmentation methods.

Keywords: Text-prompted medical image segmentation · Segment anything model 2.

1 Introduction

The Segment Anything Model (SAM) [10] has revolutionized image segmentation by introducing a promptable framework, which supports interactive segmentation using visual prompts such as points, bounding boxes, and mask inputs. SAM-2 [13] advances the architecture with a streaming memory that stores previous prompts and predictions, enabling it to segment anything in images and videos. Several works [12,16,20] have introduced SAM series into the medical image segmentation domain. For example, MedSAM-2 [20] leverages the SAM-2 pipeline and designs a self-sorting memory bank for pseudo-video data consisting of 2D images without temporal order.

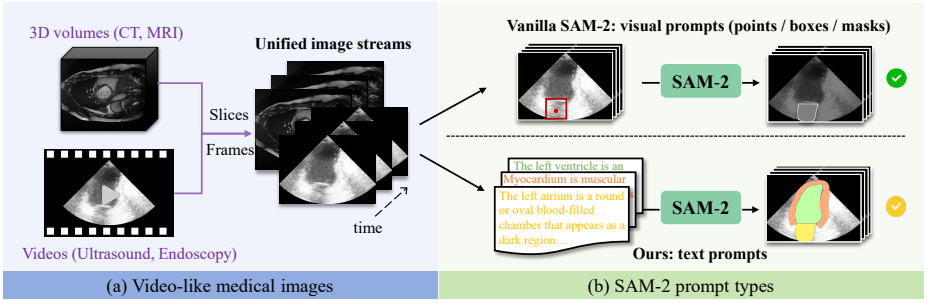

Fig. 1. Medical image segmentation using SAM-2. (a) Video-like medical images are processed into unified image streams. (b) While vanilla SAM-2 supports visual prompts, our approach incorporates text prompts for enhanced semantic understanding.

However, despite their versatility, the SAM series lack the ability to incorporate textual semantic prompts, which limits their application in scenarios requiring nuanced contextual understanding. Previous methods [11,18,4] have proved that text prompts are beneficial for improving segmentation performance via multi-modal interaction. Therefore, improving SAM-2's ability to interpret medical text prompts remains a challenge, as shown in Fig. 1 (b).

To address this gap, we propose a novel framework for **T**ext-**G**uided medical image segmentation using **S**egment **A**nything **M**odel **2** (TGSAM-2). Medical imaging modalities, such as videos (e.g., ultrasound and endoscopy) and 3D volumes (e.g., CT and MRI), share video-like properties, making them well-suited for SAM-2. These data types exhibit temporal or spatial continuity, where consecutive frames in a video or adjacent slices in a 3D volume are highly correlated, and can be processed into unified image streams, as shown in Fig. 1 (a).

Ultrasound and endoscopy data capture dynamic changes of anatomical structures or probe movements in real time. Similarly, sequential slices in CT and MRI scans can be treated as frames, in which organs maintain consistent spatial relationships but exhibit varying appearances across the entire volume. To continuously refer to target objects, we leverage text prompts that describe key attributes such as the relative position, rough color, and shape of target organs or lesions, which remain stable over time. We design a **Text-Conditioned Visual Perception (TCVP)** module to condition visual features on text prompts, along with a **Text-Tracking Memory Encoder (TTME)** to focus on target objects under textual guidance, thus enhancing memory retention. Our contributions are summarized as follows:

- We extend the SAM-2 architecture to integrate semantic understanding through text embedding, enabling text-prompted segmentation for medical images via Text-Conditioned Visual Perception.
- We propose a Text-Tracking Memory Encoder to ensure consistent target tracking across frames in dynamic medical imaging scenarios.
- We evaluate our method on diverse medical image datasets, demonstrating significant improvements over existing methods.

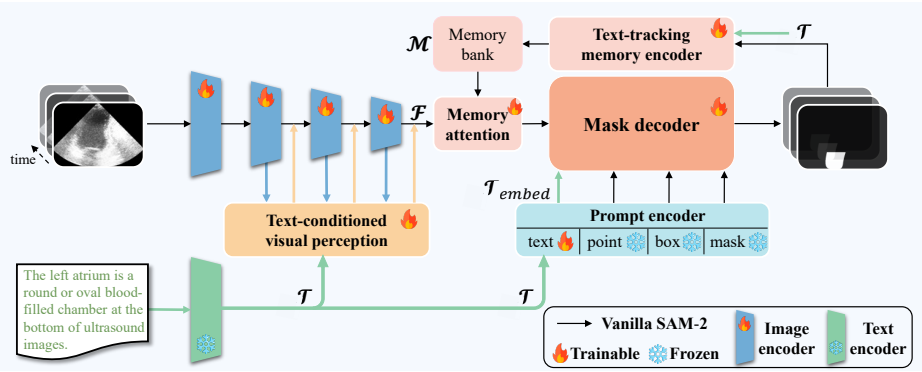

Fig. 2. Overview of TGSAM-2. Text-conditioned visual perception (TCVP) conditions visual features on text embeddings, and Text-tracking memory encoder (TTME) ensures continuous tracking of target objects under guidance of the provided text.

2 Method

The overall architecture of our TGSAM-2 is shown in Fig. 2. Given T frames $\{I_t \in \mathbb{R}^{3\times H\times W}\}_{t=1}^{T}$ from a video or volume, and their corresponding text prompt P, the model aims to predict binary mask outputs $Y = \{\hat{y}_t \in \mathbb{R}^{H\times W}\}_{t=1}^{T}$ for each object. To explain how we enhance SAM-2 with textual semantic understanding for video-like medical images, we first introduce some preliminaries and describe the process of prompting SAM-2 with medical text prompts in Section 2.1. We then discuss the role of text in guiding visual perception in Section 2.2. Finally, to maintain consistent tracking of the target object, we propose a refined memory encoder that incorporates textual features, as detailed in Section 2.3.

2.1 Overview

Preliminaries of the Segment Anything Model 2 (SAM-2) SAM-2 consists of an image encoder E_{image}, a prompt encoder E_{prompt}, and a mask decoder D. In addition, a memory attention A, a memory encoder E_{memory}, and a memory bank which stores memories $\mathcal{M}$ are introduced to enable video processing.

The image encoder E_{image} is an MAE [7] pre-trained Hiera [14], which is hierarchical to extract multi-scale features $\mathcal{F}$. The memory attention module A is a stack of transformer blocks that takes the current frame features and attends to the historical features $\mathcal{M}$ stored in the memory bank. The prompt encoder E_{prompt} takes visual prompts P_t (points, boxes, or masks) as input and outputs the prompt embedding $E_{prompt}(P_t)$.

The mask decoder D gets $E_{prompt}(P_t)$ from the prompt encoder and frame embeddings after memory attention A, and predicts a binary mask $\hat{y}_t$:

$$\hat{y}_t = D\left(A\left(E_{image}(I_t), \mathcal{M}\right), E_{prompt}(P_t)\right) \tag{1}$$

The output mask $\hat{y}_t$ is downsampled and fed into the memory encoder E_{memory} with the original frame features $E_{image}(I_t)$ to generate a memory. The memory bank retains memories of the past K frames for the target object in the video:

$$\mathcal{M} = \{E_{memory}(\hat{y}_t, E_{image}(I_t))\}_{t-K+1}^{t} \tag{2}$$

Prompting SAM-2 with Text To inject textual semantics into SAM-2, we adopt a text encoder E_{text} to extract textual features $\mathcal{T} \in \mathbb{R}^{L\times C}$ from the text prompt consisting of L words. In the prompt encoder E_{prompt}, these features $\mathcal{T}$ are linearly projected using a learnable projection layer $\mathcal{T}_{proj} = W \cdot \mathcal{T}$, $W \in \mathbb{R}^{C\times D}$, and then token-level information is aggregated into $\mathcal{T}_{embed} \in \mathbb{R}^{L\times D}$ via attention-based integration, which is formulated as:

$$\mathcal{T}_{embed} = \text{Softmax}\left(\frac{W_Q\mathcal{T}_{proj} \cdot (W_K\mathcal{T}_{proj})^\top}{\sqrt{D}}\right) \cdot \mathcal{T}_{proj} \tag{3}$$

We consider text as a type of sparse prompts like points and boxes, thus $\mathcal{T}_{embed}$ will be summed with positional encodings and fed into the mask decoder D.

2.2 Text-conditioned Visual Perception (TCVP)

The text encoder E_{text} extracts textual features $\mathcal{T}$, and the image encoder E_{image} extracts multi-scale features $\mathcal{F} = \{f_i \in \mathbb{R}^{C_i\times H_i\times W_i}\}_{i=1}^{N}$, where C_i, H_i and W_i denote the channel dimension, height, and width of the feature map at the i^{th} level, respectively, and N is the number of feature levels. TCVP is a multi-modal feature fusion mechanism designed to enhance visual understanding under textual guidance. This module leverages multi-head cross attention to align and integrate textual features with multi-scale visual features, enabling the model to adaptively highlight contextually relevant regions based on semantic cues from the text. As shown in Fig. 3 (a), the process can be formulated as:

$$\begin{gathered} f_N \mathrel{+}= \text{MHCA}(\mathcal{T}, f_N) \\ f_{N-1} \mathrel{+}= \text{Act}(\text{DeConv}(f_N)), f_{N-2} \mathrel{+}= \text{DeConv}(f_{N-1}) \end{gathered} \tag{4}$$

where MHCA$(\cdot)$ represents multi-head cross attention, with textual features $\mathcal{T}$ as query, and visual features of the last level f_N as key and value. DeConv$(\cdot)$ is a transposed convolutional layer, and Act$(\cdot)$ denotes the GELU activation function. Through multi-modal interaction and hierarchical feature integration, TCVP enables SAM-2 to obtain context-aware visual features for subsequent memory attention, seamlessly bridging the gap between language and vision.

2.3 Text-tracking Memory Encoder (TTME)

In medical videos, visual cues alone are often insufficient for tracking objects due to low contrast and blurry boundaries. Therefore, we design a Text-Tracking

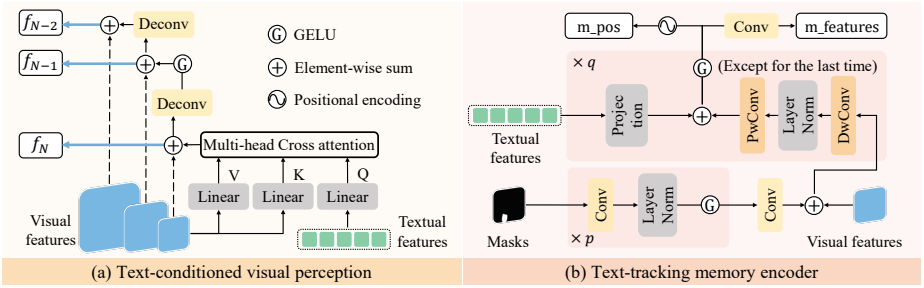

Fig. 3. Schematizations of (a) TCVP and (b) TTME.

Memory Encoder to enhance target tracking performance by incorporating textual information alongside visual features, as illustrated in Fig. 3 (b). Initially, the memory encoder E_{memory} relies on the last level of previous frames' visual features f_N and associated masks $\hat{y}$. Taking a previous frame I_s as an example, its predicted mask $\hat{y_s}$ is downsampled through several convolutional layers:

$$\hat{y}'_s = \text{Act}(\text{LN}(\text{Conv}(\hat{y_s}))), \text{repeat } p \text{ times} \tag{5}$$

Then the downsampled mask $\hat{y}'_s$ is added to the visual features $f_{N,s}$, i.e., $f'_{N,s} = \text{Conv}(\hat{y'_s}) + f_{N,s}$. We introduce textual features $\mathcal{T}$ as an additional input to our TTME for tracking objects over time. The memory feature of I_s is computed as:

$$\mathcal{M}_s = \text{Act}(\text{PwConv}(\text{LN}(\text{DwConv}(f'_{N,s}))) + W\mathcal{T}), \text{repeat } q \text{ times} \tag{6}$$

where $\text{LN}(\cdot)$ denotes layer normalization, W represents a learnable projection layer, $\text{DwConv}(\cdot)$ and $\text{PwConv}(\cdot)$ denote depth-wise and point-wise convolutions, respectively. $\text{DwConv}(\cdot)$ remains the channel dimension unchanged, while $\text{PwConv}(\cdot)$ maps features into different channels using 1×1 kernels. By adjusting the memory updates with the text, the model is able to better differentiate between objects and adapt to dynamic changes in the scene.

3 Experiments

3.1 Datasets and Metrics

We train and evaluate our model on four datasets of different modalities from the Referring Medical Image Sequence Segmentation dataset [17], a large-scale benchmark consisting of frames from video-based examinations and slices from 3D volumes, along with medical text prompts for each anatomical structure. The four datasets are: 1) **ACDC** [3] (MRI), which contains 100 training and 50 testing volumes for segmenting left ventricle, right ventricle, and myocardium. 2) MSD **Spleen** segmentation dataset [1,15] (CT), comprising 30 training and 11 testing volumes. 3) Micro-Ultrasound **Prostate** Segmentation Dataset [9] (ultrasound), containing 55 training and 20 testing videos. 4) **CVC**-ClinicDB [2] (endoscopy), containing 18 training and 11 testing videos with polyps. The Dice score (DSC) and Intersection over Union (IoU) are used as evaluation metrics.

Table 1. Comparison results on ACDC, Spleen, Prostate, and CVC datasets. † denotes the average of left ventricle, right ventricle, and myocardium.

Method	ACDC†		Spleen		Prostate		CVC	
	DSC↑	IoU↑	DSC↑	IoU↑	DSC↑	IoU↑	DSC↑	IoU↑
Task-specific								
UNet++ (TMI19)	83.25	75.24	77.88	64.60	85.65	80.59	72.46	62.43
nn-UNet (Nature21)	86.54	81.98	86.98	81.99	89.73	83.64	80.34	72.15
TransUNet (MIA24)	86.45	81.71	87.50	82.13	88.98	83.26	77.95	69.66
Text-guided								
LViT (TMI24)	84.64	76.46	81.82	75.00	90.41	84.69	70.56	61.56
LanGuide (MICCAI23)	84.82	74.16	88.24	78.96	91.50	84.34	75.87	61.13
MMI-UNet (MICCAI24)	85.78	75.08	88.78	79.83	90.29	82.30	78.75	64.95
Interactive (point-prompted)								
MedSAM (Nature24)	85.47	79.05	89.32	84.00	90.60	85.10	79.31	73.60
SAM-2 (ArXiv24)	84.29	78.84	85.15	79.10	88.69	83.38	83.60	76.58
MedSAM-2 (ArXiv24)	86.04	79.32	87.75	81.02	91.57	86.14	84.35	77.26
TGSAM-2 (Ours)	**87.63**	**82.10**	**89.34**	**84.77**	**92.75**	**87.74**	**85.10**	**78.27**

3.2 Implementation Details

We adopt the pretrained sam2_hiera_small model with 46M parameters as the initial weights. Images are resized to 1024×1024. The size of memory bank is set to 4. We use BiomedBERT [6] as the text encoder, which is pretrained using abstracts from PubMed and full-text articles from PubMedCentral. The channel dimension C of textual features is set to 256. The repeat times p and q in text-tracking memory encoder are set to 4 and 2, respectively. Our model is trained on an RTX 3090 24GB GPU, using Adam optimizer with an initial learning rate of 1e-4. The learning rate decays by 0.5 every 10 epochs.

3.3 Main Results

Comparison with State-of-the-art Methods We compare our method with previous state-of-the-art approaches, which can be categorized into three types: 1) *Task-specific* models, including UNet++ [19], TransUNet [5], and nn-UNet [8]. 2) *Text-guided* models, including LViT [11], LanGuide [18], and MMI-UNet [4]. 3) *Interactive* models using *point prompts* every 5 frames, including MedSAM [12], SAM-2 [13], and MedSAM-2 [20]. We train separate models for each organ and lesion, as shown in Table 1. Our method demonstrates superior performance over other methods, indicating the effectiveness and generalization ability of equipping SAM-2 with medical text prompts. We also visualize segmentation predictions across different modalities in Fig. 4 (a)-(c).

Table 2. Performance on ACDC dataset. specific: Models are trained on each cardiac structures separately. universal: Models are trained on all three structures collectively. ♯: Model is prompted every 5 frame. *: Model is prompted every frame.

Method	Prompt	Type	Left Ven.		Right Ven.		Myocardium		Average	
			DSC↑	IoU↑	DSC↑	IoU↑	DSC↑	IoU↑	DSC↑	IoU↑
UNet++	-	specific	87.41	81.71	80.03	71.86	82.33	72.16	83.25	75.24
MedSAM-2♯	point	specific	89.45	83.98	82.21	74.38	86.46	79.59	86.04	79.32
Ours	text	specific	91.57	87.14	85.28	79.95	86.05	79.22	**87.63**	**82.10**
MedSAM-2*	point	universal	81.38	73.73	75.20	67.94	36.37	28.21	64.32	56.63
Ours	text	universal	90.62	85.17	81.50	74.49	84.97	75.11	85.70	78.26

Semantic-aware Performance When multiple objects are present in a single medical image, task-specific models are trained using class-specific masks, while interactive models utilize class-agnostic points (clicks). Medical text prompts, which carry semantic information, allow our method to distinguish between different objects. As shown in Table 2, we train a universal model on the ACDC dataset solely using text prompts, achieving an average DSC of 85.70% and IoU of 78.26%, which outperforms the universal MedSAM-2 model and is comparable to task-specific MedSAM-2. Visualization results are shown in Fig. 4 (d)-(f).

3.4 Ablation Studies

Component Analysis To analyze the effectiveness of the Text-conditioned Visual Perception (TCVP) and Text-tracking Memory Encoder (TTME), we conduct ablation studies, as shown in Table 3. The TCVP and TTME components improve the DSC by 2.75% and 2.18%, respectively, demonstrating that medical text prompts can assist in feature extraction and object tracking.

Prompt Design Medical text prompts used in our experiments contain descriptions of organs/lesions, including attributes such as definition, color, and shape. We evaluate the impact of these prompts, as shown in Table 4. Without

Table 3. Ablation studies on TCVP and TTME.

TCVP	TTME	ACDC		Spleen		Prostate		CVC		Average	
		DSC↑	IoU↑	DSC↑	IoU↑	DSC↑	IoU↑	DSC↑	IoU↑	DSC↑	IoU↑
-	-	84.19	78.45	85.61	79.94	88.27	83.09	80.46	72.62	84.63	78.53
✓	-	86.87	81.81	87.78	82.83	91.08	85.77	83.79	78.21	87.38	82.16
-	✓	86.12	81.09	87.20	82.11	90.85	85.45	83.05	77.80	86.81	81.61
✓	✓	**87.63**	**82.10**	**89.34**	**84.77**	**92.75**	**87.74**	**85.10**	**78.27**	**88.71**	**83.22**

Table 4. Ablation on prompt design.

Text prompt	Average	
	DSC↑	IoU↑
w/o text	84.63	78.53
class name	86.40	80.95
description	**88.71**	**83.22**

Table 5. Ablation on text-tracking.

Text-tracking	Average	
	DSC↑	IoU↑
multiplication	87.05	82.53
concatenation	85.14	79.88
summation	**88.71**	**83.22**

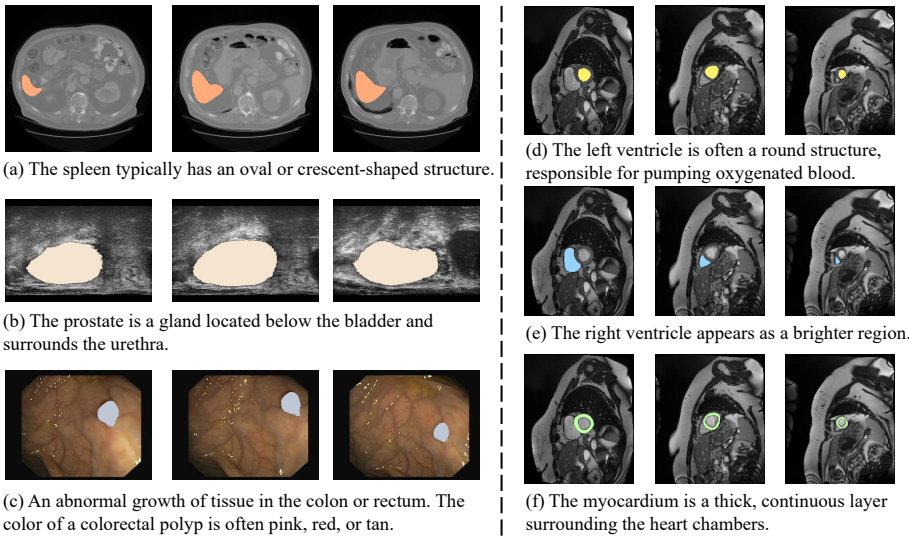

Fig. 4. Visualization of text-guided segmentation results.

such prompts, the model achieves an average DSC of 84.63% and IoU of 78.53%. Refining the prompts from simple class names to detailed descriptions improves performance by 2.31% in DSC and 2.27% in IoU.

Text-tracking Strategy As mentioned in Section 2.3 Eq. (6), we perform a summation of textual features with a combination of visual features and predicted masks. We explore different strategies to integrate them, including element-wise multiplication and channel-wise concatenation. Results in Table 5 show that summation outperforms both multiplication and concatenation, with improvements of 1.66% and 3.57% in DSC, respectively.

4 Conclusion

In this work, we present TGSAM-2 that utilizes text prompts to enhance SAM-2 for medical image segmentation. Through the integration of Text-Conditioned

Visual Perception and Text-Tracking Memory Encoder, our method demonstrates improvements across diverse datasets with different modalities. Experimental results emphasize the value of text prompts in medical imaging. In the future, we plan to extend our framework to incorporate visual prompts and unstructured non-visual cues, such as patient metadata and clinical history.

Acknowledgments. This work was supported by Shanghai Natural Science Foundation (No. 25ZR1401028), and the Science and Technology Commission of Shanghai Municipality (No. 23511100602; No. 21511104506).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. 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.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015)
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., 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)
4. Bui, P.N., Le, D.T., Choo, H.: Visual-textual matching attention for lesion segmentation in chest images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 702–711. Springer (2024)
5. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis **97**, 103280 (2024)
6. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)
7. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
9. Jiang, H., Imran, M., Muralidharan, P., Patel, A., Pensa, J., Liang, M., Benidir, T., Grajo, J.R., Joseph, J.P., Terry, R., et al.: Microsegnet: A deep learning approach for prostate segmentation on micro-ultrasound images. Computerized Medical Imaging and Graphics p. 102326 (2024)

10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
11. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE transactions on medical imaging (2023)
12. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
13. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
14. Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al.: Hiera: A hierarchical vision transformer without the bells-and-whistles. In: International Conference on Machine Learning. pp. 29441–29454. PMLR (2023)
15. Simpson, A.L., Leal, J.N., Pugalenthi, A., Allen, P.J., DeMatteo, R.P., Fong, Y., Gönen, M., Jarnagin, W.R., Kingham, T.P., Miga, M.I., et al.: Chemotherapy-induced splenic volume increase is independently associated with major complications after hepatic resection for metastatic colorectal cancer. Journal of the American College of Surgeons **220**(3), 271–280 (2015)
16. Wu, J., Ji, W., Liu, Y., Fu, H., Xu, M., Xu, Y., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
17. Yuan, R., Xu, J., Chen, M., Li, Q., Zhang, Y., Feng, R., Zhang, T., Gao, S.: Text-promptable propagation for referring medical image sequence segmentation. arXiv preprint arXiv:2502.11093 (2025)
18. Zhong, Y., Xu, M., Liang, K., Chen, K., Wu, M.: Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 724–733. Springer (2023)
19. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging **39**(6), 1856–1867 (2019)
20. Zhu, J., Qi, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)