LTSE: Language-guided Tissue Referring Segmentation in Pathology Images with Adaptive Expert Mixture

Jiao Tang[1], Bo Qian[1], Peng Wan[1], Wei Shao (✉)[1], and Daoqiang Zhang (✉)[1]

College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing 211106, China
shaowei20022005@nuaa.edu.cn, dqzhang@nuaa.edu.cn

Abstract. Tissue segmentation is essential for pathology image analysis. Conventional deep learning based segmentation methods require large amounts of annotated data and are constrained by the predefined classes, making them less flexible in adapting to diverse clinical requirements and user-specific queries. The language-guided referring segmentation (LGRS) model can help identify and segment specific objects based on user-provided descriptions. However, the existing LGRS models lack the capability to explicitly reject nonexistent targets, and struggle in effectively segmenting multiple target regions. Based on the above considerations, we propose LTSE, a language-guided tissue referring segmentation assistant for pathology images, which inherits the powerful multi-modal alignment capabilities of Multi-modal Large Language Models (MLLMs) to implement tissue segmentation according to the instructions. Specifically, we expand the original vocabulary with multiple [SEG] tokens to support multiple mask references and a [REJ] token to reject the empty targets. In addition, we enhance the adaptability and accuracy in multi-target segmentation by developing an Adaptive Expert Mixture (AEM) module that can dynamically select specialized expert decoders based on the textual and visual characteristics of the input data. We for the first time curate a vision-language pathology dataset BCSS-Ref for the tissue referring segmentation task with matched images, masks and textual information, and the experimental results demonstrate the superiority of our method in comparison with the existing studies.

Keywords: Tissue Referring Segmentation · Mixture of Experts

1 Introduction

Tissue segmentation is a crucial task in pathology image analysis, serving as a critical component in disease diagnosis, prognosis, and treatment planning [27].

Traditional deep learning based segmentation methods have made remarkable progress, with convolutional neural networks (CNNs) [21] and transformer-based models [3][2] demonstrating high accuracy in delineating tissue structures of

pathology images. However, these methods typically require extensive manually annotated data, which is costly and labor-intensive [24]. Furthermore, they are constrained by a fixed set of predefined categories, which makes it difficult to segment tissue regions based on user-specified descriptions, thereby limiting their flexibility in adapting to complex and dynamic clinical scenarios [9].

To overcome these challenges, language-guided referring segmentation (LGRS) has emerged as a promising alternative, allowing more adaptive and flexible target identification through language instructions [8]. Previous LGRS studies mainly focused on fusing images and language to segment objects [6][28]. For instance, Li *et al.* [15] developed a new text-augmented medical image segmentation model LViT on X-rays and CT images. Ouyang *et al.* [22] proposed a language-guided scale-aware medical segmentor LSMS to segment various liver lesions. In order to bridge the modality gap for better performance, most of the existing studies employed cross-attention mechanisms or cross-modal alignment techniques to effectively integrate visual and linguistic features [5][4][16]. Based on the rapid development of Multi-modal Large Language Models (MLLMs) that can efficiently align the vision and language modalities with extraordinary performance, Lai *et al.* [13] presented a language-instructed segmentation assistant LISA for referring and reasoning segmentation on natural images. Xia *et al.* [26] introduced a vision segmentation assistant GSVA for further performance improvement. Other studies include [19][30] presented a generalized decoding framework that can predict pixel-level segmentation with language instructions.

Although much progress has been achieved, most existing LGRS methods lack the ability to reject irrelevant or empty targets. As a matter of fact, different tissues in the complex tumor micro-environment might appear quite alike in terms of their overall structure and cellular arrangement. For instance, the endothelial cells and tumor cells usually show similar appearance during the vessel formation process [12]. Accordingly, there is a high possibility that the endothelial cells may be mis-segmented as tumor cells if the target of tumor cells are not rejected. In addition, the existing methods struggle with segmenting multiple target regions effectively as they rely on a unified decoder, which limits their ability to adapt to the diverse and complex language instructions.

Based on the above considerations, we propose LTSE, a novel language-guided tissue referring segmentation assistant, which inherits the powerful multi-modal alignment capabilities of MLLMs to generate precise segmentation masks based on the language instructions. Specifically, we expand the original vocabulary of the LLM with multiple [SEG] tokens to support multi-target identification and a [REJ] token to explicitly reject empty targets. Instead of relying on a unified mask decoder, we further propose an Adaptive Expert Mixture (AEM) module to dynamically select specialized expert decoders that can enhance the model's adaptability to diverse and complex language instructions. We for the first time curate a vision-language pathology dataset BCSS-Ref for tissue referring segmentation task with matched images, masks and textual information. The experimental results demonstrate the superiority of our method in comparison with the existing studies.

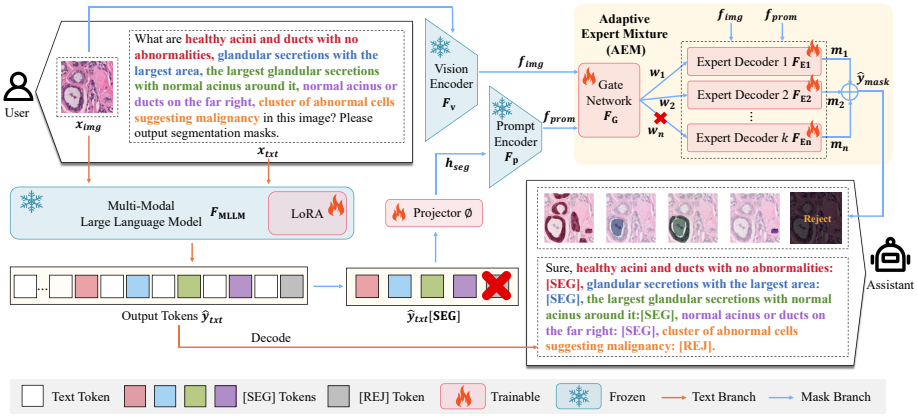

Fig. 1. Overview of LTSE, which consists of text branch and mask branch. In the text branch, given an input image and its corresponding text instructions, the MLLM generates output tokens which are then decoded as text output. LTSE generates multiple [SEG] tokens for multiple referred regions and a [REJ] token to reject empty targets. In the mask branch, all [SEG] tokens are selected to prompt the mask decoding process. The Adaptive Expert Mixture (AEM) module assigns weights to expert decoders via a gate network and discards irrelevant experts, and then mixes their outputs to segment the target objects referred to the instructions.

2 Method

The architecture of LTSE consists of two branches: text branch and mask branch. We show the flowchart of our method in Fig. 1.

2.1 Text Branch: Generation of Text Responses

In the text branch, we combine the input text instructions and corresponding images to generate the text responses. Specifically, given a text instruction $\boldsymbol{x}_{txt}$ along with the input image $\boldsymbol{x}_{img}$, we feed them into the Multi-modal Large Language Model (MLLM) F_{MLLM} to derive the output tokens $\hat{\boldsymbol{y}}_{txt}$:

$$\hat{\boldsymbol{y}}_{txt} = F_{\text{MLLM}}(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}). \tag{1}$$

Then, text responses are generated from $\hat{\boldsymbol{y}}_{txt}$ using a linear classifier to predict next words in the vocabulary. Notably, LTSE supports up to M [SEG] tokens for multiple target regions and a [REJ] token to reject empty targets in $\hat{\boldsymbol{y}}_{txt}$.

2.2 Mask Branch: Segmentation with Adaptive Expert Decoders

The mask branch of LTSE focuses on generating segmentation masks for target regions based on the [SEG] tokens and image features using an Adaptive Expert Mixture (AEM), where each expert serves as a mask decoder. Specifically, we

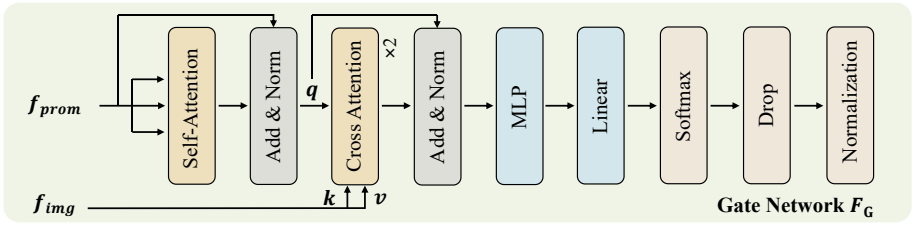

Fig. 2. Architecture of the gate network.

first encode the input image $\boldsymbol{x}_{img}$ by a vision encoder F_{v} to extract the image features $\boldsymbol{f}_{img}$ for segmentation:

$$\boldsymbol{f}_{img} = F_{\mathrm{v}}(\boldsymbol{x}_{img}). \tag{2}$$

Then, we select the output embeddings of the [SEG] tokens as well as discarding the [REJ] token from $\hat{\boldsymbol{y}}_{txt}$ to derive $\hat{\boldsymbol{y}}_{txt}$[SEG], and project it into the prompt space using an MLP projector ϕ with output $\boldsymbol{h}_{seg}$. Next, we pass $\boldsymbol{h}_{seg}$ through a prompt encoder F_{p} to obtain the prompt embedding $\boldsymbol{f}_{prom}$:

$$\boldsymbol{h}_{seg} = \phi(\hat{\boldsymbol{y}}_{txt}[\mathrm{SEG}]), \boldsymbol{f}_{prom} = F_{\mathrm{p}}(\boldsymbol{h}_{seg}). \tag{3}$$

Finally, instead of relying on a unified mask decoder for tissue segmentation that is not adaptive to the complex text instructions, we design an Adaptive Expert Mixture (AEM) module to dynamically select specialized expert decoders for enhancing the segmentation adaptability and accuracy. Specifically, suppose that we have n expert decoders $\{F_{\mathrm{E1}}, F_{\mathrm{E2}}, \cdots, F_{\mathrm{En}}\}$ where the structure of each individual decoder follows [11]. We denote the output of the decoders as $\boldsymbol{m} = [\boldsymbol{m}_1, \boldsymbol{m}_2, \cdots, \boldsymbol{m}_n]$, where

$$\boldsymbol{m}_i = F_{\mathrm{Ei}}(\boldsymbol{f}_{img}, \boldsymbol{f}_{prom}). \tag{4}$$

The AEM assigns weights $\boldsymbol{w} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_n]$ to each expert decoder via a gate network F_{G}:

$$\boldsymbol{w} = F_{\mathrm{G}}(\boldsymbol{f}_{img}, \boldsymbol{f}_{prom}). \tag{5}$$

Fig. 2 presents the detailed architecture of the gate network F_{G}. Specifically, we first apply self-attention to $\boldsymbol{f}_{prom}$ to capture the textual information. Then, two layers of cross-attention are performed, using $\boldsymbol{f}_{prom}$ as the query and $\boldsymbol{f}_{img}$ as the key and value to integrate both textual and visual features. After each attention layer, addition and normalization operations are applied to stabilize the learning process. The output is then processed using MLP and linear mapping, followed by a softmax operation to compute the weights for each expert decoder. To improve the model efficiency and focus on the most relevant experts, we retain the top k experts based on their weights, setting the rest to zero. The final weights $\boldsymbol{w}$ are then obtained through normalization. To obtain the segmentation

masks $\hat{\boldsymbol{y}}_{mask}$ for target regions, we mix the outputs of expert decoders according to their weights:

$$\hat{\boldsymbol{y}}_{mask} = \sum_{i=1}^{n} \boldsymbol{w}_i \cdot \boldsymbol{m}_i. \tag{6}$$

2.3 Training Objectives

Our model is trained by simultaneously optimizing the text generation loss $\mathcal{L}_{txt}$ and the segmentation mask loss $\mathcal{L}_{mask}$. The overall objective $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \mathcal{L}_{txt} + \mathcal{L}_{mask}, \tag{7}$$

where $\mathcal{L}_{txt}$ is the auto-regressive cross-entropy loss [20] for text generation, $\mathcal{L}_{mask}$ is the combination of per-pixel binary cross-entropy (BCE) loss [14] and DICE loss [29]:

$$\begin{aligned} \mathcal{L}_{txt} &= \mathbf{CE}(\hat{\boldsymbol{y}}_{txt}, \boldsymbol{y}_{txt}), \\ \mathcal{L}_{mask} &= \mathbf{BCE}(\hat{\boldsymbol{y}}_{mask}, \boldsymbol{y}_{mask}) + \mathbf{DICE}(\hat{\boldsymbol{y}}_{mask}, \boldsymbol{y}_{mask}), \end{aligned} \tag{8}$$

where $\boldsymbol{y}_{txt}$ and $\boldsymbol{y}_{mask}$ represent ground-truth text responses and segmentation masks, respectively.

3 Experiment

Implementation Details. We use LLaVA-7B-v1-1 [17] as the base Multimodal Large Language Model (MLLM) that is fine-tuned with LoRA [7]. In addition, we adopt ViT-H SAM [11] backbone as the vision encoder F_{v} and the prompt encoder F_{p}. The number of expert decoders n is set to 4, and $k = 2$ experts are retained in AEM. LTSE supports up to $M = 10$ [SEG] tokens, enabling it to handle multiple referred regions. We adopt 4 NVIDIA 24G 4090 GPUs for training, and the training scripts are based on deepspeed engine [23]. We use AdamW optimizer and set the learning rate as 3e-4 without weight decay. We use WarmupDecayLR as the learning rate scheduler, allocating 100 iterations for warmup. Additionally, the batch size is set to 2 per device, with a gradient accumulation step of 10. We train LTSE for 10 epochs, with 500 steps per epoch.

Dataset. We for the first time curate a vision-language pathology dataset BCSS-Ref for tissue referring segmentation task with matched images, masks and textual information. Our BCSS-Ref is based on the BCSS dataset [1], which includes 151 whole-slide images (WSIs). We invite pathology experts to review the BCSS semantic annotations firstly, and then provide detailed descriptions on the important regions in each WSI. We randomly split the dataset into 5 folds, with 4 folds (121 WSIs) used for training and the remaining (30 WSIs) used for performance evaluation. The WSIs with detailed descriptions are divided

Table 1. Comparison results of tissue referring segmentation on BCSS-Ref dataset. '-' indicates that the method fails to reject empty targets, resulting in an N-Acc of 0.

Method	20x			40x			Overall		
	gIoU	cIoU	N-Acc	gIoU	cIoU	N-Acc	gIoU	cIoU	N-Acc
Conventional LGRS Methods									
LTS [10]	50.21	46.72	-	69.98	51.89	-	60.10	49.31	-
VLT [5]	49.78	46.89	-	70.51	52.02	-	60.15	49.46	-
CRIS [25]	51.43	48.01	-	71.18	53.70	-	61.31	50.86	-
LAVT [28]	53.72	50.07	-	73.65	55.13	-	63.69	52.60	-
ReLA [16]	61.63	54.17	91.82	79.09	61.02	89.58	70.36	57.60	90.70
X-Decoder [30]	53.91	52.15	-	74.01	57.21	-	63.96	54.68	-
LViT [15]	51.02	47.18	-	70.11	52.07	-	60.57	49.63	-
PloyFormer [18]	55.12	53.17	-	75.09	60.87	-	65.11	57.02	-
SEEM [31]	54.03	52.98	-	74.23	58.18	-	64.13	55.58	-
LSMS [22]	52.96	49.92	-	71.68	54.47	-	62.32	52.20	-
MLLM-based LGRS Methods									
LISA [13]	54.40	55.66	-	74.98	61.50	-	64.69	58.58	-
GSVA [26]	63.71	54.28	94.84	80.89	61.53	95.93	72.30	57.91	95.39
LTSE (ours)	**65.63**	**56.99**	**98.22**	**82.60**	**62.55**	**98.38**	**74.12**	**59.77**	**98.30**

into patches at 20x magnification (8,357 patches) and 40x magnification (36,651 patches). Additionally, we extend the region annotations on the patches by including their positions and area information. Our dataset link is listed below: https://pan.baidu.com/s/15pnneQRXO6TTnJL1QHxjTQ?pwd=wudj.

Evaluation Metrics. We follow the study in [26] that adopts gIoU, cIoU to evaluate the performance for referring segmentation. Specifically, gIoU computes the average IoU for each mask, while cIoU calculates the cumulative intersection area relative to the cumulative union area across the entire dataset. To evaluate empty targets, we use the measurement of No-target Accuracy (N-Acc) [26], which measures the ratio of correctly classified empty-target text expressions to the total number of empty-target text expressions in the dataset.

Comparison with the State-of-the-Arts. We compare LTSE with the following 12 methods: **LTS** (CVPR21) [10], **VLT** (ICCV21) [5], **CRIS** (CVPR22) [25], **LAVT** (CVPR22) [28], **ReLA** (CVPR23) [16], **X-Decoder** (CVPR23) [30], **LViT** (TMI23) [15], **PolyFormer** (CVPR23) [18], **SEEM** (NIPS24) [31], **LSMS** [22], **LISA** (CVPR24) [13], **GSVA** (CVPR24) [26]. Among these methods, only LViT [15] and LSMS [22] are specifically designed for referring segmentation on medical images. Meanwhile, only ReLA [16] and GSVA [26] has the ability to reject empty targets. In Table 1, we show the referring segmentation results of our method and its competitors on the BCSS-Ref dataset for tissue segmentation at both 20x and 40x magnification levels. It is obvious that LTSE achieves the best segmentation results in comparison with all its competitors. Specifically, LTSE, GSVA [26] and LISA [13] are based on the MLLM that can more effectively align image and text data than the traditional LGRS methods

Table 2. Ablation study of tissue referring segmentation on BCSS-Ref dataset. '-' indicates that the variant fails to reject empty targets, resulting in an N-Acc of 0. 'M-[SEG] Tokens' indicates Multiple [SEG] Tokens.

M-[SEG] Tokens	[REJ] Token	20x			40x			Overall		
		gIoU	cIoU	N-Acc	gIoU	cIoU	N-Acc	gIoU	cIoU	N-Acc
✓	✗	59.13	55.03	-	78.87	61.28	-	69.00	58.16	-
✗	✓	50.43	49.69	29.65	69.13	57.54	32.04	59.78	53.62	30.85
✗	✗	54.41	54.96	-	74.92	60.75	-	64.67	57.86	-
LTSE		**65.63**	**56.99**	**98.22**	**82.60**	**62.55**	**98.38**	**74.12**	**59.77**	**98.30**
w/o AEM		63.19	54.70	98.05	79.96	60.60	98.19	71.58	57.65	98.12

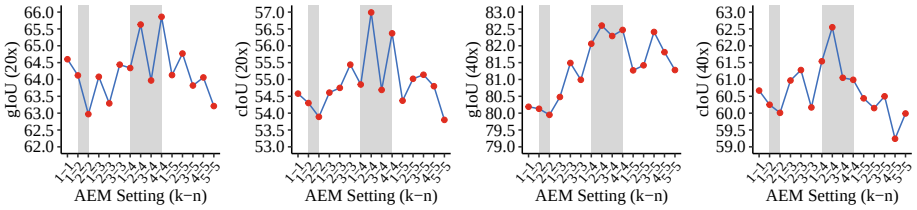

Fig. 3. Comparison Results of different AEM settings.

[10][5][25][28][16][30][15][18][31][22]. In addition, in contrast to LISA [13], LTSE and GSVA [26] show their advantages for multi-target segmentation by utilizing multiple [SEG] tokens. Moreover, our LTSE also achieves higher segmentation accuracy than GSVA [26] since it employs an Adaptive Expert Mixture (AEM) module for mask decoding that can dynamically select specialized decoders. In other words, AEM provides more flexibility and adaptability to diverse instructions and thus can lead to an improvement of 2% on the indexes of gIoU and cIoU over GSVA [26]. Finally, LTSE still outperforms ReLA [16] and GSVA [26] in rejecting empty targets, demonstrating its effectiveness in handling empty-target scenarios.

Ablation Study. To further evaluate the effectiveness of LTSE, we compare it with its variants in Table 2. First, we investigate the impact of using multiple [SEG] tokens and the [REJ] token. Here, the variant without applying the multiple [SEG] tokens means that only one [SEG] token is employed for tissue referring segmentation. As shown in Table 2, on one hand, the absence of the [REJ] token will lead to the failure of rejecting empty targets. On the other hand, solely relying on single [SEG] token will lead to a 15% drop in gIoU (shown in the 2nd row) as it struggles to segment multiple target regions. Moreover, LTSE also outperforms its variant without the AEM module (shown in the last row), highlighting its advantage in enhancing segmentation performance by selectively activating relevant expert decoders.

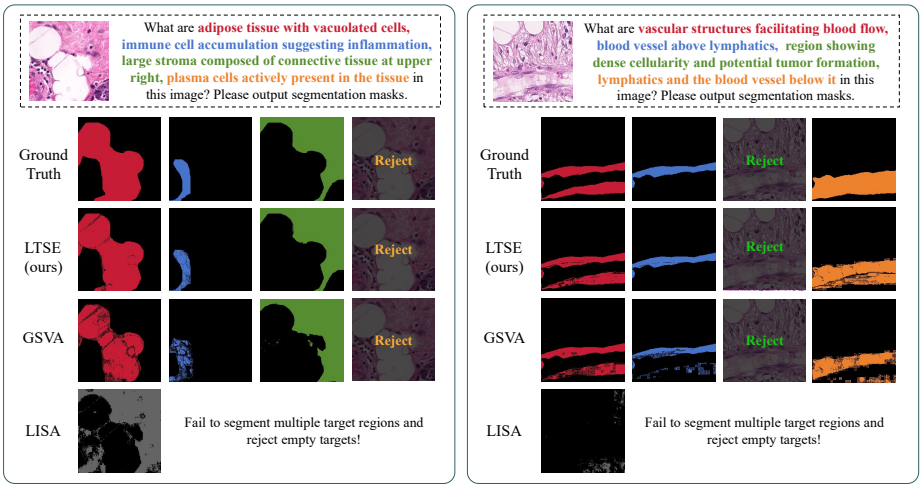

Fig. 4. Comparison of visualization results among different methods.

Discussion of AEM. In Fig. 3, we discuss the impact of different AEM settings, where k-n denotes the use of n expert decoders with k decoders are activated. As shown in Fig. 3, the segmentation accuracy will be significantly improved when the expert decoder number (n) increases from 1 to 4, but will drop when n reaches to 5. The reason lies in the fact that adding more experts will strengthen the model's ability to handle diverse instructions. However, the involvement of too many experts will introduce unnecessary complexity and may bring noise without improving the segmentation results. Furthermore, it is also easy to observe that the suitable number of activated experts (k) can lead to higher segmentation results than simply choosing the best decoder ($k = 1$) or considering all decoders (k=n), which validates the advantage of taking both diversity and individual capability of expert decoders into consideration for referring segmentation.

Visualization Results. Fig. 4 presents the visualization results for different methods. On one hand, as a representative LGRS method, LISA [13] struggles with multi-target segmentation and empty-target rejection, while our LTSE effectively handles these cases. On the other hand, LTSE also outperforms the current SoTA method GSVA [26] for tissue segmentation since it can provide more consistent segmentation results.

4 Conclusion

Tissue referring segmentation plays a vital role in clinical pathology, enabling precise identification and delineation of regions of interest in pathology images based on textual descriptions. In this paper, we for the first time curate a vision-language pathology dataset BCSS-Ref with matched images, masks and textual

information, and propose a novel language-guided tissue referring segmentation assistant LTSE for multi-target segmentation as well as rejecting empty targets. Our LTSE also involves an Adaptive Expert Mixture (AEM) module that can dynamically select relevant expert decoders to improve segmentation accuracy. Experimental results on the BCSS-Ref dataset verify the potential of our LTSE for referring segmenting on complex tumor micro-environment components.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Nos. 62136004, 62272226, 62102188), Key Research and Development Plan of Jiangsu Province, China under Grant BE2022842, Postdoctoral Fellowship Program of CPSF (No. GZC20242233).

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

References

1. Amgad, M., Elfandy, H., Hussein, H., Atteya, L.A., Elsebaie, M.A., Abo Elnasr, L.S., Sakr, R.A., Salem, H.S., Ismail, A.F., Saad, A.M., et al.: Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics **35**(18), 3461–3467 (2019)
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Chng, Y.X., Zheng, H., Han, Y., Qiu, X., Huang, G.: Mask grounding for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26573–26583 (2024)
5. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16321–16330 (2021)
6. Feng, G., Hu, Z., Zhang, L., Lu, H.: Encoder fusion network with co-attention embedding for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15506–15515 (2021)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
8. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 108–124. Springer (2016)
9. Jayasinghe, S.: Describing complex clinical scenarios at the bed-side: Is a systems science approach useful? exploring a novel diagrammatic approach to facilitate clinical reasoning. BMC medical education **16**, 1–6 (2016)
10. Jing, Y., Kong, T., Wang, W., Wang, L., Li, L., Tan, T.: Locate then segment: A strong pipeline for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9858–9867 (2021)

11. 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)
12. Kramer, R.H., Nicolson, G.L.: Interactions of tumor cells with vascular endothelial cell monolayers: a model for metastatic invasion. Proceedings of the National Academy of Sciences **76**(11), 5704–5708 (2009)
13. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024)
14. Li, Q., Jia, X., Zhou, J., Shen, L., Duan, J.: Rediscovering bce loss for uniform classification. arXiv preprint arXiv:2403.07289 (2024)
15. 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)
16. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23592–23601 (2023)
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
18. Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R.K., Mahadevan, V., Manmatha, R.: Polyformer: Referring image segmentation as sequential polygon generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18653–18663 (2023)
19. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55. Springer (2024)
20. Mao, A., Mohri, M., Zhong, Y.: Cross-entropy loss functions: Theoretical analysis and applications. In: International conference on Machine learning. pp. 23803–23828. PMLR (2023)
21. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
22. Ouyang, S., Zhang, J., Lin, X., Wang, X., Chen, Q., Chen, Y.W., Lin, L.: Lsms: Language-guided scale-aware medsegmentor for medical image referring segmentation. arXiv preprint arXiv:2408.17347 (2024)
23. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3505–3506 (2020)
24. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Medical Image Analysis **67**, 101813 (2021)
25. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022)
26. Xia, Z., Han, D., Han, Y., Pan, X., Song, S., Huang, G.: Gsva: Generalized segmentation via multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3858–3869 (2024)

27. Xing, F., Yang, L.: Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. IEEE reviews in biomedical engineering **9**, 234–263 (2016)
28. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165 (2022)
29. Zhao, R., Qian, B., Zhang, X., Li, Y., Wei, R., Liu, Y., Pan, Y.: Rethinking dice loss for medical image segmentation. In: 2020 IEEE International Conference on Data Mining (ICDM). pp. 851–860. IEEE (2020)
30. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (2023)
31. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36** (2024)