Evaluation and Improvement of Segment Anything Model for interactive histopathology image segmentation

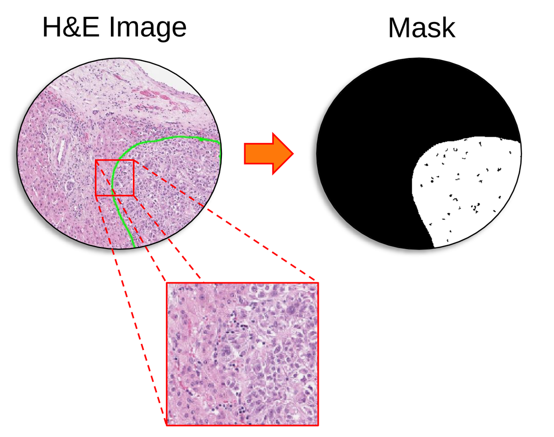
SeungKyu Kim, Hyun-Jic Oh, Seonghui Min and Won-Ki Jeong*

Korea University, College of Informatics,
Department of Computer Science and Engineering

PAPER ↓

Introduction

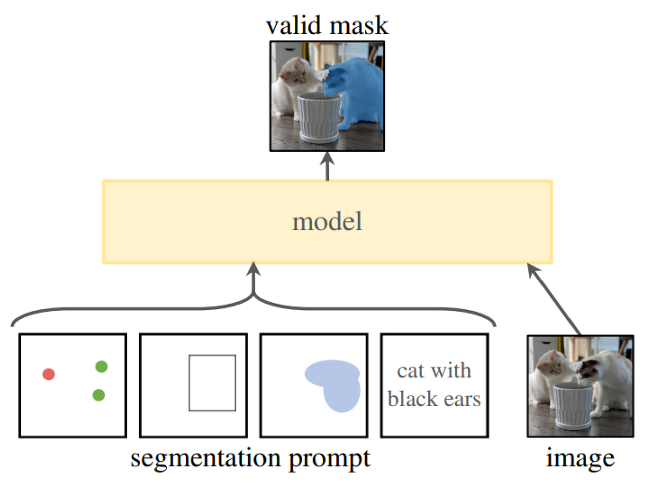
Task & Problem definition



Tumor region segmentation in WSIs

- **WSIs: Indistinct** and **ambiguous** boundaries.
- **General fully-supervised** approaches require **extensive** and **accurately annotated datasets.**
- Weakness in **generalization** ability.

SAM [1]



- **Promptable foundation model.**
- **Trained on huge natural dataset** (SA-1B) which includes 11million images and over 1 billion masks.
- Consists of Image Encoder, Prompt Encoder and Mask Decoder.

Motivation

① **Could we directly use SAM for histopathology?**

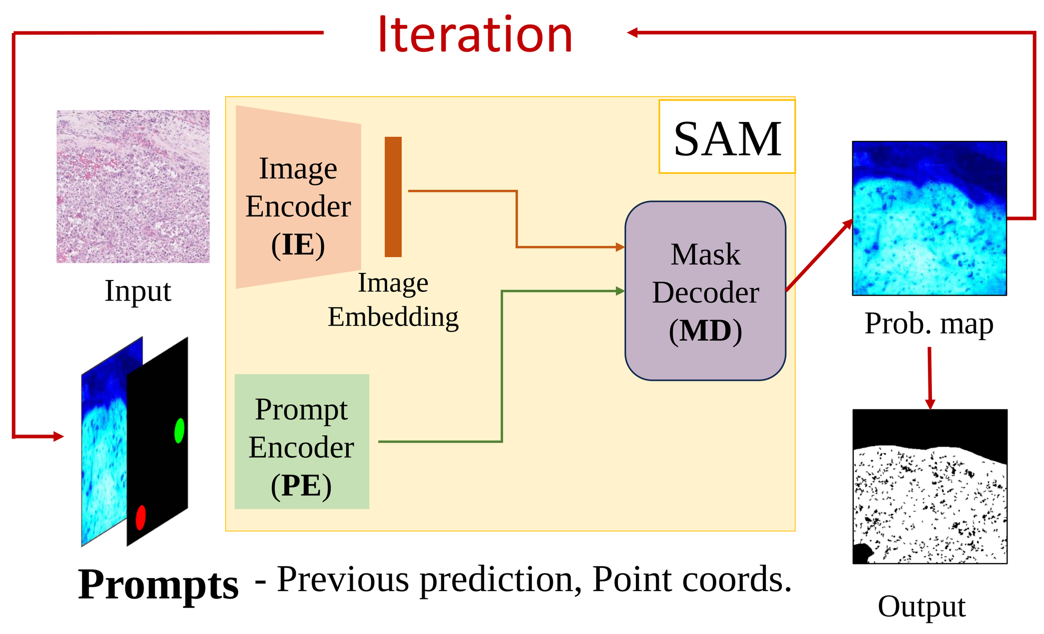
② **If not, how to efficiently utilize it?**

Contributions

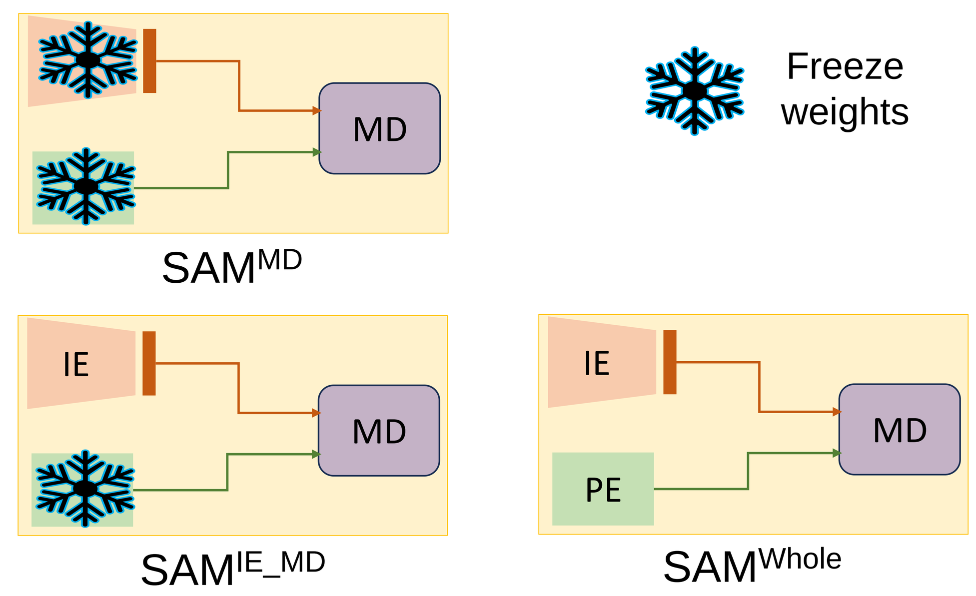
- **Assessed SAM's capability for Zero-Shot** histopathology image segmentation in the context of interactive segmentation by comparing it against SOTA interactive methods[2-4].
- Explored various fine-tuning scenarios for SAM, **providing insights into their effective utilization**.
- Introduced an **enhanced mask decoder**, reducing fine-tuning costs while preserving SAM's strength.

Methods

① Interactive Segmentation using SAM

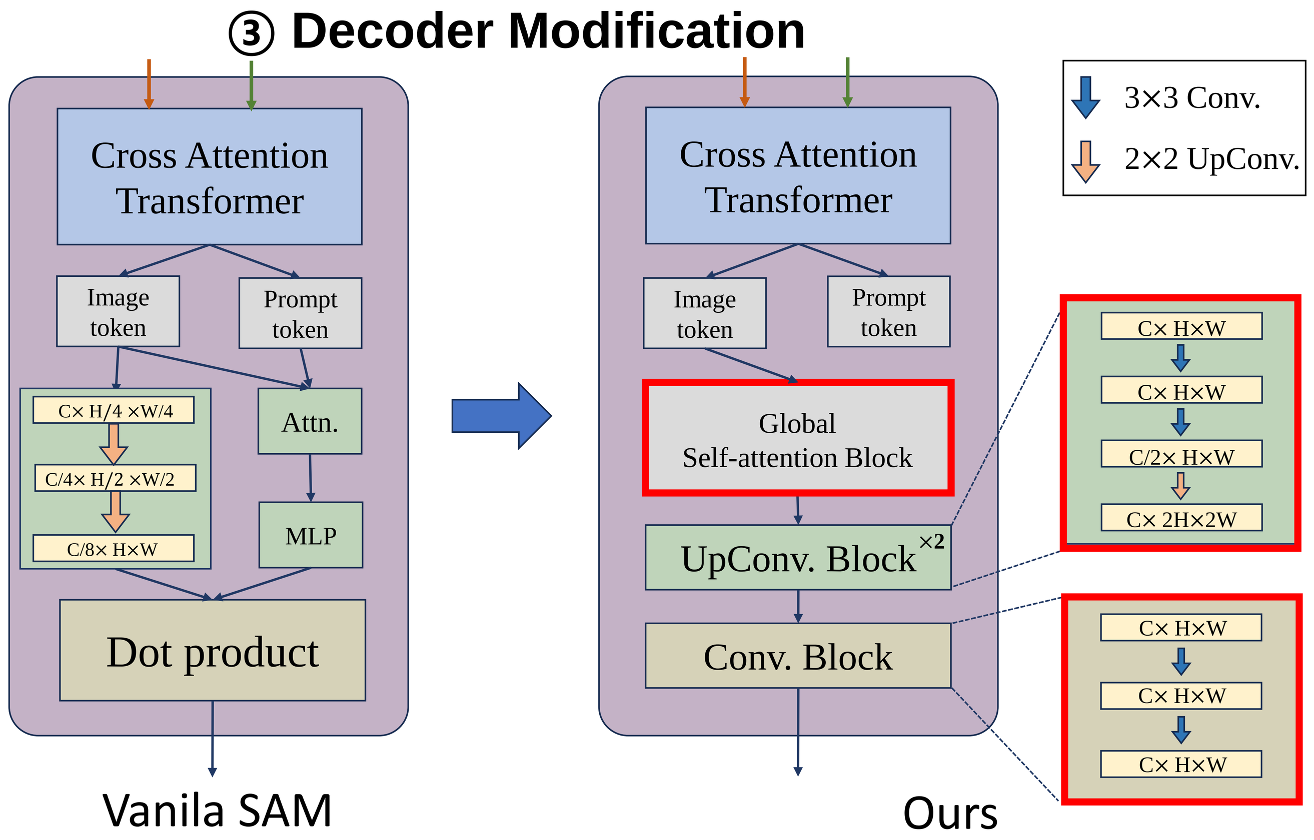


② Three fine-tuning scenarios



- For understanding impact of each component

③ Decoder Modification



Global attention block: For capturing global context.

Deepen layers: For increasing representational capacity of the decoder.

★ Only mask decoder is trained
→ 8x reduction in training cost compared to training whole model

Experiment results

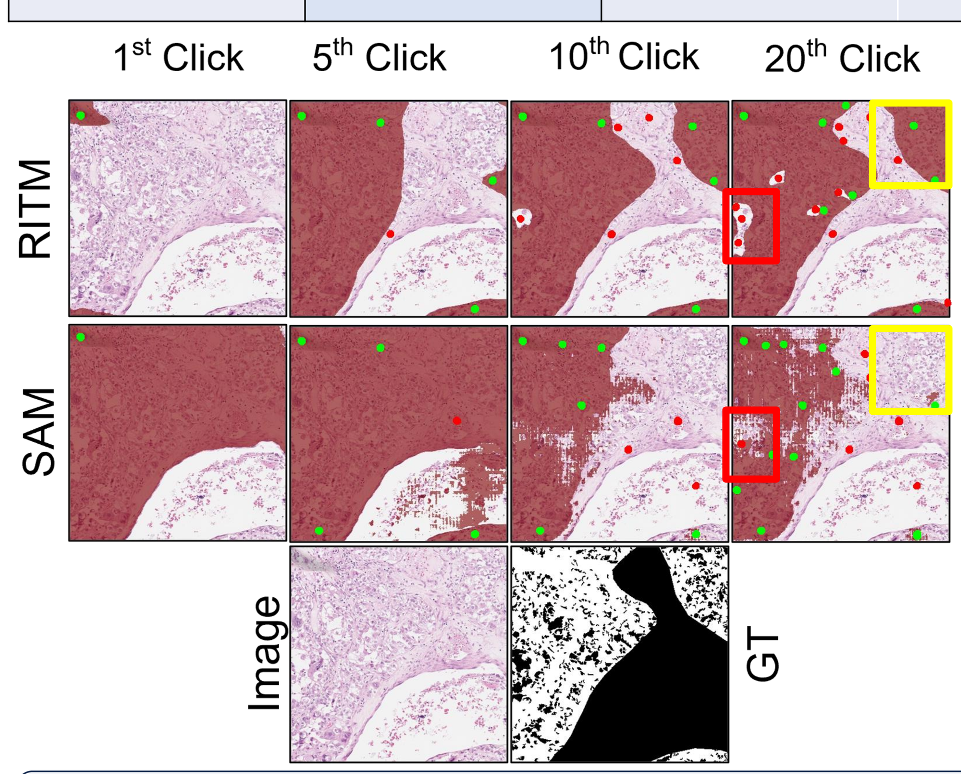
① Zero-Shot performance: SOTA methods vs. SAM

Dataset	Method	NoC@ (↓) 80	NoC@ (↓) 85	NoC@ (↓) 90	SPC(s) (↓)
PAIP 2019 (x5)	RITM	7.43	10.24	13.00	0.075
	Focal	7.37	**9.89**	**12.42**	0.073
	Simple	**7.21**	10.16	13.44	0.189
	SAM	9.13	12.10	14.73	**0.052**
CAMELYON16 (x10)	RITM	6.91	8.19	10.07	0.077
	Focal	**4.73**	**5.89**	**7.87**	0.076
	Simple	5.20	6.42	8.55	0.187
	SAM	6.64	8.49	11.03	**0.053**

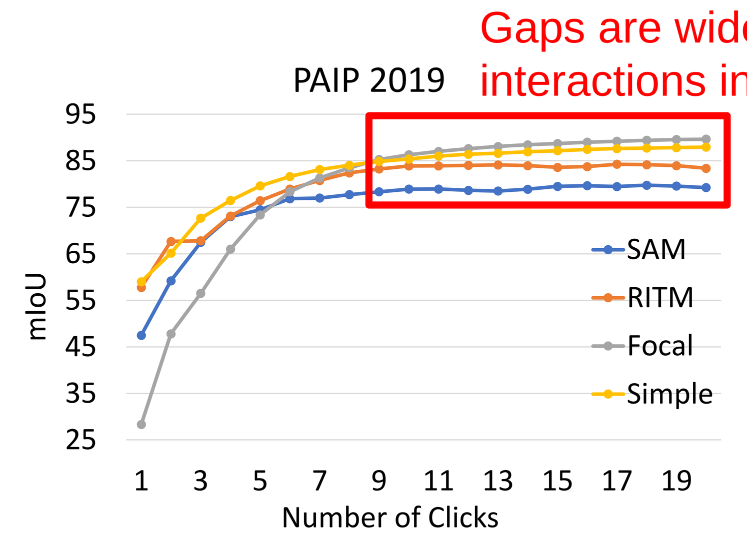
Accuracy is lower but faster !

② After fine-tuning on PAIP2019: SOTA methods / 3 scenarios / Ours

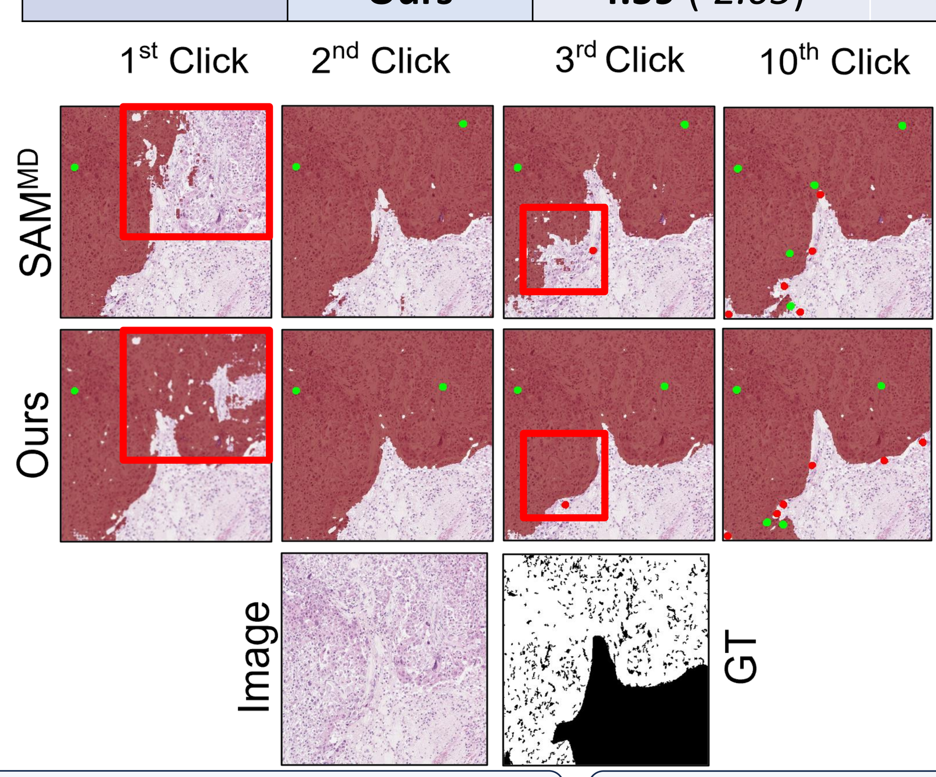
Dataset	Method	NoC@ (↓) 80	NoC@ (↓) 85	NoC@ (↓) 90	SPC(s) (↓)
PAIP 2019 (x5)	RITM	<u>2.78</u>	4.93	9.40	0.075
	Focal	**2.58**	**4.54**	**8.98**	0.071
	Simple	5.11	8.33	12.20	0.189
	SAM^{MD}	5.80	8.93	12.09	**0.050**
	SAM^{IE_MD}	4.53	7.58	10.86	**0.050**
	SAM^{Whole}	4.53	7.50	10.95	**0.052**
	Ours	4.75	7.78	10.85	<u>0.067</u>
CAMELYON16 (x10)	RITM	6.28 (-0.63)	7.65 (-0.54)	9.67 (-0.31)	0.076
	Focal	11.82 (+7.09)	13.01 (+7.12)	14.53 (+7.33)	0.076
	Simple	6.07 (+0.87)	7.44 (+1.02)	9.94 (+1.39)	0.187
	SAM^{MD}	<u>4.88</u> (-1.76)	6.68 (-1.81)	9.07 (-1.96)	**0.053**
	SAM^{IE_MD}	7.63 (+0.99)	9.19 (+0.7)	11.81 (+0.78)	**0.049**
	SAM^{Whole}	6.60 (-0.04)	8.10 (-0.39)	10.66 (-0.37)	**0.053**
	Ours	**4.59** (-2.05)	**5.92** (-2.57)	**8.40** (-2.63)	<u>0.064</u>



Gaps are widening as interactions increase!



- Weakness in local refinement capability.
- No longer improvement of accuracy after a certain number of clicks.



- Trained solely on PAIP for assessing generalization capability.
- The values in parentheses indicate the change relative to zero-shot.
- 'Ours' shows considerable enhancement compared to SAM^{MD} which is trained under identical conditions.
- Comparable to SAM^{whole} with significantly lower training costs.
- The inference time has increased but is still faster compared to other SOTA methods.

Conclusion & Limitations

- SAM shows **strengths in generalization capability** and notably excelled in terms of **inference speed**, however, exhibits relatively lower performance compared to SOTA interactive models.
- By modifying architecture of mask decoder, we could enhance the performance while maintaining high generalization capability and fast inference speed.
- In the **PAIP dataset**, our model still exhibits a **noticeable gap** compared to SOTA models.
- The **table does not clearly highlight the strengths** of our model in an intuitive manner (Need additional quantitative metric).

References

[1] 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. arXiv preprint arXiv:2304.02643 (2023)
[2] Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., Zhao, H.: Focalclick: towards practical interactive image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1300–1309 (2022)
[3] Liu, Q., Xu, Z., Bertasius, G., Niethammer, M.: Simpleclick: Interactive image segmentation with simple vision transformers (2023)
[4] Sofiiuk, K., Petrov, I.A., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3141–3145. IEEE (2022)

Acknowledgement

This work was partially supported by the National Research Foundation of Korea (NRF-2019M3E5D2A01063819, NRF-2021R1A6A1A 13044830), the Institute for Information & Communications Technology Planning & Evaluation (IITP-2023-2020-0-01819), the Korea Health Industry Development Institute (HI18C0316), the Korea Institute of Science and Technology (KIST) Institutional Program (2E32210 and 2E32211), and a Korea University Grant.

High-Performance Visual Computing Lab, Korea University
ksk8804@korea.ac.kr | wkjeong@korea.ac.kr | hvcl.korea.ac.kr

2023 1st International Workshop on Foundation Models for General Medical AI