Zero-shot capability of 2D SAM-family models for bone segmentation in CT scans

Caroline Magg[1,2][0009−0008−5592−2586], Hoel Kervadec[1,2][0000−0002−6786−7042], and Clara I. Sánchez[1,2][0000−0001−9787−8319]

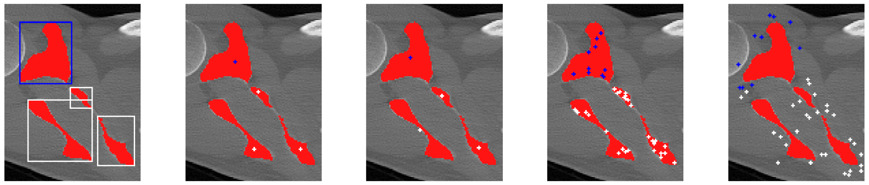
[1] University of Amsterdam, The Netherlands
[2] Amsterdam UMC location University of Amsterdam, The Netherlands

1 Introduction

The Segment Anything Model (*SAM*) [5] is a promptable 2D foundation model (FM) for segmentation, recently extended to *SAM2* [10]. Evaluation studies of *SAM* and *SAM2* on various medical datasets employing different prompting strategies have shown some preliminary promising results but also limitations [1, 3, 4, 8, 9], motivating the development of FM dedicated to medical image segmentation [2,7,12,14]. To be potentially used in a clinical setting, extensive evaluation studies are crucial, to identify their weaknesses and risks; while developing guidelines for robust and efficient prompting. Currently, there exists no such independent study on bone segmentation in CT scans, which, due to their distinct boundaries, should lead to promising results [9]. Thus, we perform an evaluation study to test the zero-shot capability of SAM-family models with different non-iterative prompting strategies for bone CT segmentation, on a private yet clinically representative dataset.

2 Method

Models We investigate in total 9 models of the SAM-family: *SAM* [5] (base (B), large (L), huge (H) model sizes), *SAM2* [10] (tiny (T), small (S), base+ (B+), large (L)) , *Med-SAM* [7] (unique size) and *SAM-Med2D* [2] (unique size).
Data As SAM derivatives (e.g., *Med-SAM*) are fined-tuned on publicly available datasets [6, 11, 13]), we have to resort to a private (Amsterdam UMC) dataset we compiled, to ensure a fair comparison across all models. It contains two different skeletal regions: 15 bilateral shoulder CT scans with labels for scapula and humerus and 40 unilateral wrist CT scans with labels for capitate, lunate, radius, scaphoid, triquetrum, and ulna.
Prompting strategies We use non-iterative prompts, automatically extracted from reference masks. There are 5 prompt primitives (`bounding box`, `center`, `centroid`, `positive` and `negative` points) and two variants (`single` and `multiple`), see Fig. 1. The primitives are used either individually (`bbox`, `center`, `centroid`, 1/3/5/10 `positive` points) or in combinations (`bbox` + `center`, `bbox` + 1 or 5 `positive` or `negative` points) in both variants, for a total of 24 strategies per model; except *Med-SAM* which only supports bounding boxes.
Metrics The Dice Coefficient (Dice) and 95%-percentile Hausdorff distance (HD95) are used as performance metrics.

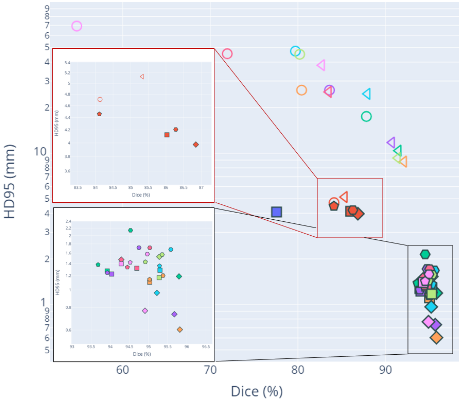


(a) `bbox` (b) `center` (c) `centroid` (d) `positive` (e) `negative`

Fig. 1: Prompt primitives: (a) bounding box, (b) center, (c) centroid, (d) positive random points inside the object, (e) negative random points outside the object. The largest component's prompt is blue (i.e., `single`), while the others are white, resulting in the `multiple` setting when all prompts are used.

3 Results

A prompt subset, based on their overall ranking of Dice performance, is displayed in Fig. 2. The best results—all involving the `bbox` primitive—are located in the lower right corner of Fig. 2, i.e., high Dice and low HD95.



	`bbox multiple`	`bbox + center mult.`	`bbox + pos. mult.`	`bbox + neg. mult.`	`center sin.`	`1 random sin.`
Models	■	◆	⬢	⬟	○	◁
● Med-SAM	$77.6_{15.8}$	–	–	–	–	–
● SAM-Med2d	$86.0_{7.6}$	$\mathbf{86.8}_{8.0}$	$86.3_{7.9}$	$84.1_{8.4}$	$84.1_{8.7}$	$85.3_{8.7}$
● SAM B	$93.9_{3.5}$	$\mathbf{95.8}_{3.5}$	$94.5_{3.5}$	$93.7_{3.6}$	$87.8_{9.2}$	$91.5_{7.3}$
● SAM L	$94.0_{3.2}$	$\mathbf{95.7}_{4.1}$	$94.7_{3.6}$	$93.9_{3.2}$	$83.6_{11.1}$	$90.7_{7.4}$
● SAM H	$95.0_{2.8}$	$\mathbf{95.8}_{3.0}$	$95.4_{3.0}$	$95.0_{2.6}$	$80.4_{10.8}$	$92.1_{6.7}$
● SAM2 T	$95.3_{3.3}$	$95.2_{5.1}$	$\mathbf{95.6}_{3.8}$	$95.3_{3.2}$	$79.7_{15.7}$	$87.9_{12.2}$
● SAM2 S	$94.7_{3.3}$	$94.3_{6.4}$	$\mathbf{95.0}_{3.2}$	$94.4_{3.3}$	$71.9_{17.3}$	$83.5_{14.2}$
● SAM2 B+	$95.3_{3.0}$	$95.3_{4.8}$	$\mathbf{95.4}_{2.8}$	$94.9_{3.1}$	$80.2_{13.7}$	$91.4_{6.8}$
● SAM2 L	$94.3_{4.1}$	$94.9_{5.5}$	$\mathbf{95.0}_{4.1}$	$94.5_{3.8}$	$54.8_{22.4}$	$82.7_{18.0}$

Read: Average $\text{DSC}_{\text{standard deviation}}$

Fig. 2: (Left) Performance scatter plot, (right) Dice (%) performance of best settings.

4 Conclusion

Testing the zero-shot capability of SAM-family models on bone CT scans is crucial to assess risks for clinical application and define an optimal prompting strategy. Despite the size of our current dataset, valuable conclusions can already be drawn: the medically fine-tuned SAM-versions are outperformed by the original *SAM* and *SAM2*. Point prompts—such as `center`, `centroid` or `positive`—independent of the number of points, show higher HD95 and lower Dice as prompts involving only bounding boxes, whereas bounding box combined with center points shows the best performances. We plan to extend our dataset and perform a more exhaustive evaluation of the models in follow-up works.

Acknowledgments. We thank in alphabetical order Leendert Blankevoort, Johannes G.G. Dobbe, Geert J. Streekstra, Stela Topalova, and Lukas P.E. Verweij for their work in data curation.

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

References

1. Cheng, D., Qin, Z., Jiang, Z., Zhang, S., Lao, Q., Li, K.: Sam on medical images: A comprehensive study on three prompt modes (2023), https://arxiv.org/abs/2305.00035
2. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Sun, L.J.H., He, J., Zhang, S., Zhu, M., Qiao, Y.: Sam-med2d (2023)
3. Dong, H., Gu, H., Chen, Y., Yang, J., Chen, Y., Mazurowski, M.A.: Segment anything model 2: an application to 2d and 3d medical images (2024), https://arxiv.org/abs/2408.00756
4. He, S., Bao, R., Li, J., Stout, J., Bjornerud, A., Grant, P.E., Ou, Y.: Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets (2023), https://arxiv.org/abs/2304.09324
5. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023), https://arxiv.org/abs/2304.02643
6. Liu, P., Han, H., Du, Y., Zhu, H., Li, Y., Gu, F., Xiao, H., Li, J., Zhao, C., Xiao, L., Wu, X., Zhou, S.K.: Deep learning to segment pelvic bones: Large-scale ct datasets and baseline models (2021), https://arxiv.org/abs/2012.08721
7. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 1–9 (2024)
8. Mattjie, C., De Moura, L.V., Ravazio, R., Kupssinskü, L., Parraga, O., Delucis, M.M., Barros, R.C.: Zero-shot performance of the segment anything model (sam) in 2d medical imaging: A comprehensive evaluation and practical guidelines. In: 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE). pp. 108–112 (2023). https://doi.org/10.1109/BIBE60311.2023.00025
9. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: An experimental study. Medical Image Analysis **89**, 102918 (2023). https://doi.org/https://doi.org/10.1016/j.media.2023.102918, https://www.sciencedirect.com/science/article/pii/S1361841523001780
10. 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), https://arxiv.org/abs/2408.00714
11. Sekuboyina, A., Husseini, M.E., Bayat, A., Löffler, M., Liebl, H., Li, H., Tetteh, G., Kukačka, J., Payer, C., Štern, D., Urschler, M., Chen, M., Cheng, D., Lessmann, N., Hu, Y., Wang, T., Yang, D., Xu, D., Ambellan, F., Amiranashvili, T., Ehlke, M., Lamecker, H., Lehnert, S., Lirio, M., de Olaguer, N.P., Ramm, H., Sahu, M., Tack, A., Zachow, S., Jiang, T., Ma, X., Angerman, C., Wang, X., Brown, K., Kirszenberg, A., Élodie Puybareau, Chen, D., Bai, Y., Rapazzo, B.H., Yeah, T., Zhang, A., Xu, S., Hou, F., He, Z., Zeng, C., Xiangshang, Z., Liming, X., Netherton, T.J., Mumme, R.P., Court, L.E., Huang, Z., He, C., Wang, L.W., Ling, S.H., Huỳnh, L.D., Boutry, N., Jakubicek, R., Chmelik, J., Mulay, S., Sivaprakasam, M., Paetzold, J.C., Shit, S., Ezhov, I., Wiestler, B., Glocker,

B., Valentinitsch, A., Rempfler, M., Menze, B.H., Kirschke, J.S.: Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images. Medical Image Analysis **73**, 102166 (2021). https://doi.org/https://doi.org/10.1016/j.media.2021.102166, https://www.sciencedirect.com/science/article/pii/S1361841521002127

12. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., Qiao, Y.: Sam-med3d: Towards general-purpose segmentation models for volumetric medical images (2024), https://arxiv.org/abs/2310.15161
13. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023). https://doi.org/10.1148/ryai.230024, https://doi.org/10.1148/ryai.230024
14. Zhu, J., Qi, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2 (2024), https://arxiv.org/abs/2408.00874