Medical SAM for LGE-MRI cardiac segmentation: promise or hype ?

Celia Goujat, Pierre-Marc Jodoin, Loic Boussel, Olivier Bernard

Medical SAM for LGE-MRI cardiac segmentation: promise or hype?

Celia Goujat[1], Pierre-Marc Jodoin[3], Loïc Boussel[1,2], and Olivier Bernard[1]

[1] INSA-Lyon, Universite Claude Bernard Lyon 1, CNRS, Inserm, CREATIS UMR 5220, U1294, F-42023, Saint Etienne, France
celia.goujat@insa-lyon.fr
[2] Department of Radiology, Hospices Civils de Lyon, Lyon, France
[3] Department of Computer Science, University of Sherbrooke, Sherbrooke, QC, Canada

Abstract. Late gadolinium enhancement (LGE) cardiac magnetic resonance (MR) imaging is a key technique for assessing myocardial damage in various cardiovascular diseases. Precise identification of cardiac structures is essential for computing clinical biomarkers used in diagnosis and prognosis. While many deep learning methods have been proposed for automatic segmentation, their generalizability is hindered by limited training data and domain adaptation issues, restricting clinical deployment. Foundation models have recently emerged as a promising solution, able to perform zero and few-shot segmentation across various imaging modalities through large-scale pretraining. Among them, the Segment Anything Model (SAM) has become a widely recognized reference, with adaptations for medical imaging, such as MedSAM and SAM-Med2D.
In this study, we evaluate MedSAM and SamMed2D for segmenting the left ventricle and myocardium in LGE MR images from a private dataset comprising 135 patients. We first demonstrate that zero-shot performance remains limited, due to the scarcity of LGE MR data in their pretraining. Next, we show that fine-tuning the MedSAM decoder significantly improves segmentation quality, surpassing the nnU-Net baseline, though it requires precise bounding box initialization. We thus propose a modified MedSAM architecture that enables multi-class segmentation from a single bounding box. However, our experiments reveal that despite various improvements, MedSAM continues to produce mixed results. While our approach can segment multiple structures with one BB, it still requires accurate initialization, and its performance converges towards that achieved by nnU-Net.

Keywords: Foundation model · Cardiac imaging · Segmentation.

1 Introduction

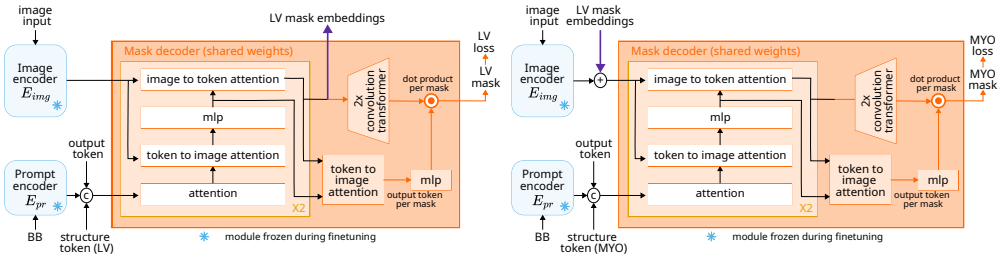
The application of foundation models for medical image segmentation has seen in rapid expansion in recent years, driven by the advent of the SAM model [9].

These methods are semi-automatic and require initialization using prompts such as bounding box (BB), points, masks, or text. MedSAM [10] and SAM-Med2D [3] are among the most advanced models for medical image segmentation. MedSAM is built upon the original SAM architecture and was trained with a large-scale medical image dataset comprising 1.6 M image-mask pairs, covering 10 different imaging modalities. This large-scale dataset allows the model to learn a rich representation of medical images, capturing a broad spectrum of anatomies and lesions across different modalities. MedSAM fine-tunes both the encoder and decoder of SAM and contains 94 million parameters. As for SAM-Med2D, it was also adapted from SAM but kept the original prompting system which includes BB, points, and masks, allowing for more diverse initializations. SAM-Med2D was trained on a larger dataset composed of 4.6M images and 19.7M masks from public and private datasets, comprising various modalities and objects. Also, given that adapters have proven effective for fine-tuning large-scale models [1], the authors adopted this approach to train the SAM encoder on domain-specific medical imaging tasks, while preserving the model's original knowledge. Overall, SAM-Med2D contains a total of 271 million parameters.

Compared to classical approaches such as nnU-Net [8], which are well established in medical image segmentation, medical SAM models appear to be better suited for handling domain adaptation issues. However, the approach also presents notable limitations, including: (1) the reliance on a prompting system, which makes the method inherently semi-automatic, and (2) the binary nature of the output for each prompt, which restricts it to single-class segmentation. While the former has been widely studied, the latter remains underexplored in the literature. Indeed, early studies have highlighted the sensitivity of Medical SAM models to both the type [13,6] and quality [2,11,12] of the input prompts. Consequently, several studies aimed to automate prompt generation, with the goal of achieving fully automatic segmentation [7,15]. While these methods demonstrate the feasibility of eliminating manual prompt initialization, they also tend to show a drop in segmentation accuracy, which limits their practical utility.

In this study, we conduct a comprehensive evaluation of medical SAM models for the segmentation of cardiac structures in LGE MR images. Although these models have been trained on large-scale datasets that include some cardiac data, the representation of LGE MR images remains limited. To assess their generalization capabilities under realistic conditions, we curated a dedicated private dataset composed of 135 patients. Our main contributions are as follows:

- We benchmarked MedSAM and SamMed2D in both zero-shot and fine-tuning scenarios, comparing their results against a baseline nnU-Net model known for its accuracy on LGE MR images [5].
- We designed targeted experiments to rigorously assess the impact of BB precision on the segmentation quality of the left ventricle (LV) and myocardium (MYO), during both training and inference phases.
- We introduced *mcMedSAM*, a novel SAM-based architecture for multi-class segmentation that enables consistent delineation of both the LV and MYO from a single BB initialization, without the need for any post-processing.



(a) LV segmentation process (b) MYO segmentation process

Fig. 1: Proposed mcMedSAM architecture. During training, the segmentation processes for the LV and MYO are applied successively at each iteration to update the shared weights of the mask decoder. During inference, the decoder is invoked twice under different input conditions to produce coherent segmentation masks for both the LV and MYO.

2 Methodology

Medical SAM models face two key limitations: the requirement for an accurate BB initialization and an inherent restriction to binary segmentation. In this work, we introduce *mcMedSAM*, a medical SAM-based architecture designed to overcome both challenges by extending the SAM framework to support multi-class segmentation from a single BB prompt. mcMedSAM relies on a high-capacity image encoder (E_{img}) to capture the rich and complex features of medical images. A prompt encoder (E_{pr}), enhanced with positional encodings [14], embeds user-provided initialization input, specifically the BB representations. These embeddings are then fused by a mask decoder (D_{mask}), which generates a segmentation mask. As will be discussed later, the same decoder is used to generate both LV and MYO masks under varying input conditions. Based on preliminary experiments presented in Section 3, we selected the MedSAM architecture as the foundation for integrating our multi-class segmentation strategy. An overview of the proposed architecture is shown in Figure 1.

Structure token - mcMedSAM contains a structure token that enables the segmentation of multiple structures given a single BB. This token is a one-hot vector representing the anatomical class to be segmented which is then passed through a learnable linear projection layer. The resulting embedding is then concatenated with the output of the prompt encoder E_{pr} (bottom left of Fig. 1-a and b). This simple yet effective mechanism allows the network to distinguish which structure—either the LV or the MYO—should be segmented from the same BB input.

Multi-class architecture - We designed a strategy that leverages the ability of MedSAM to incorporate multiple sources of information as prompts, with the

goal of exploiting anatomical relationships between cardiac structures—particularly the fact that the MYO surrounds the LV and shares a common boundary with it. Like MedSAM, the input image is first processed by E_{img} to generate an image feature map. A BB enclosing the MYO is passed to E_{pr} resulting into a sparse prompt embedding. Contrary to conventional medical SAM models, our approach processes the same mask decoder in two successive steps to optimize segmentation masks for both the LV and MYO.

In the first step (Fig. 1-a), the decoder is trained to segment the LV using only the BB surrounding the MYO and the image feature map as inputs. To inform the decoder that the target structure is the LV, a structure-specific token corresponding to the LV is concatenated with the prompt embedding and the output token, which is then used to generate the final segmentation mask. During this step, the output of the cross-attention mechanism is stored to construct a mask embedding for the LV.

In the second step (Fig. 1-b), the same decoder is further fine-tuned to segment the MYO using the same BB prompt and the previously saved LV mask embedding. The goal is to provide the decoder with additional semantic context, thereby improving the consistency between the predicted anatomical structures. As in the first step, the prompt embedding is concatenated with the structure token—now representing the MYO—and the output token. A key difference, however, is that the LV mask embedding is added to the image feature map to inject complementary semantic information.

During inference, the same decoder is invoked twice using the same input BB, but under different conditions—specifically, distinct structure tokens and the optional use of the LV mask embedding—to coherently segment both the LV and the MYO.

Decoupled version of the proposed architecture - Inspired by the work of [4], we developed a decoupled version of our method by integrating a Prompt-Decoupled Mask Module (PDMM) and applying minor modifications to the mask decoder, following the design principles outlined in [4] and illustrated in Figure 1 of that paper. The PDMM fuses the cross attention output of the prompt embeddings and feature map with multi-scale image embeddings from the encoder to reduce the over-reliance on the prompt. The corresponding architectural changes are highlighted by the dashed lines in Figure 2 of this paper. For additional implementation details, we refer the reader to the original publication.

3 Experiments

3.1 Dataset and implementation details

To evaluate our method, we used a private dataset of LGE MR images from 135 patients not involved in the pretraining of the Medical SAM models. In accordance with French regulations, this retrospective study was approved by the local ethics committee (Scientific and Ethical Committee of the Hospices

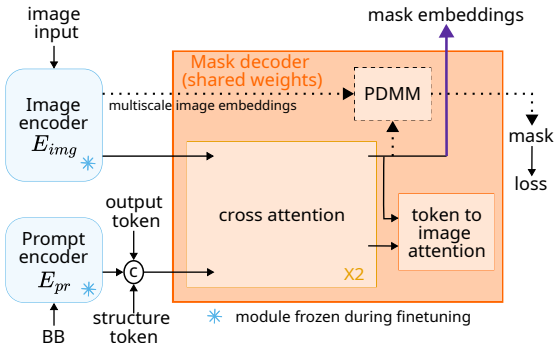


Fig. 2: Modified decoder for the decoupled mcMedSAM architecture. The dashed lines indicate the modifications introduced to the original design. This diagram illustrates the decoder configuration used for the LV segmentation process; the same modifications are applied during the MYO segmentation process.

Civils de Lyon, France). We performed 5-fold cross-validation, using one fold as the test set and the remaining folds for training and validation in each iteration. The reported results correspond to the average performance across the five folds.

Segmentation performance was evaluated using the Dice score and the 2D Hausdorff distance between manual contours and predicted masks for both the LV and MYO. For MedSAM-based models, the pretrained weights were loaded and to leverage the large-scale pretraining of MedSAM, the parameters of both the image encoder E_{img} and the prompt encoder E_{pr} were frozen. To improve computational efficiency, each image was processed once with E_{img}, and the resulting embeddings z_i were cached for reuse during both training and testing. This strategy significantly reduces training time, as only the decoder components—comprising approximately 4 million parameters—are optimized.

Models were trained using a batch size of 2 and a learning rate of 1e-5 with decay, over 50 epochs. The model achieving the best performance on the validation set was retained, ensuring convergence while minimizing overfitting. All experiments were conducted on a single Quadro RTX 8000. The loss function was a combination of Dice loss and binary cross-entropy loss.

A 2D nnU-Net model was used as the baseline for all experiments. Since Medical SAM models require a BB prompt, they were applied only to LGE MR slices that contained the anatomical structures of interest. To mitigate bias in comparisons, the nnU-Net was also trained and evaluated exclusively on the same subset of slices—i.e., those containing the relevant cardiac structures—ensuring a fair and consistent evaluation protocol across methods.

Post-processing was applied to all predicted masks using a standard approach, filling small holes within each anatomical structure as well as gaps between cardiac structures. This procedure ensured anatomically consistent and smoother segmentations, improving overall visual and quantitative quality of the results.

3.2 What about the generalization ability of medical SAM models?

We first evaluated MedSAM and SAM-Med2D in both zero-shot and fine-tuning settings. For MedSAM, fine-tuning involved updating only the decoder parameters, whereas for SAM-Med2D, both the adapter encoder and decoder were optimized. During fine-tuning, it is possible to introduce random variations in BB positioning to reduce sensitivity to initialization errors at inference. This aspect was explored in detail in the following section. In the present experiments, we used perfectly aligned BBs for both training and inference. The results are reported in Table 1. In the zero-shot configuration, both MedSAM and SAM-Med2D achieved satisfactory performance on the LV, albeit lower than that of nnU-Net, and failed to accurately segment the MYO. This limitation may stem from the fact that annular-like shapes, which characterize the MYO in LGE MR images, were likely underrepresented during the pretraining of Medical SAM models. When fine-tuned on the training set of our private dataset, both models showed improved performance, although to varying degrees. While SAM-Med2D exhibited only modest gains, MedSAM achieved remarkable results for both the LV and MYO, outperforming nnU-Net. For instance, the HD reached 2.1 mm for the LV and 3.2 mm for the MYO. This performance gap may be attributed to the fact that SAM-Med2D has approximately three times more parameters than MedSAM, making it more susceptible to overfitting and harder to fine-tune effectively on a small dataset. In addition, training time differed substantially across methods, with nnU-Net requiring on average 15h05m per fold, compared to only 2h47m for SAM-Med2D and 1h50m for MedSAM, highlighting their efficiency. This significant reduction in MedSAM training time results from our strategy of passing all images through the decoder once at the start, thereby avoiding repeated computations at each epoch. In contrast, at inference time, all images must still be processed by the encoder, which explains why MedSAM requires about 22 seconds per volume, compared to only 1.5 seconds for SAM-Med2D and 2.5 seconds for nnU-Net. Based on these results, we chose to focus the remainder of the study on the MedSAM architecture.

3.3 How sensitive is performance to BB placement accuracy?

In this section, we investigate the sensitivity of the fine-tuned MedSAM model to inaccuracies in BB positioning during both training and inference. To this end, BBs were generated from the reference masks and randomly perturbed along the horizontal and vertical axes by shifts uniformly sampled within the range of 0 to ± 6 mm. Each model was fine-tuned using this strategy and subsequently evaluated under the same conditions, with BB perturbations of comparable magnitude applied during inference. The resulting performance was compared to that of nnU-Net, with the corresponding scores presented in Figure 3. For clarity, we report the mean values averaged over the LV and MYO structures. These results highlight the importance of aligning the BB variations introduced during fine-tuning with those expected at inference. Specifically, the model trained without BB variation achieved a Dice score of approximately 95% when evaluated under

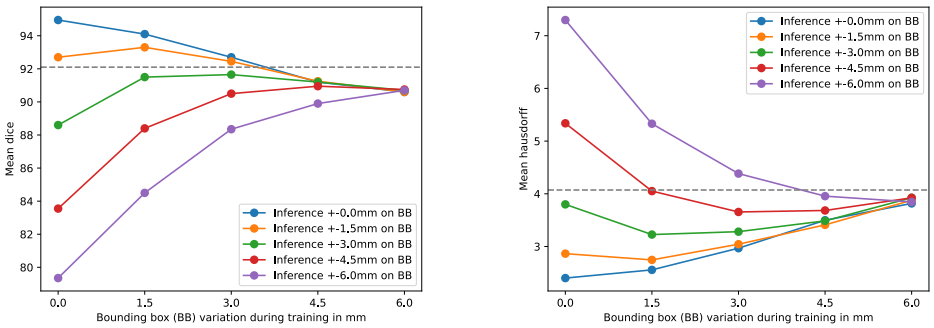
Table 1: Performance comparison between medical SAM models and the nnU-Net baseline on a private LGE MR dataset.

Method	**LV**		**MYO**		**Training time**
	Dice(%)	HD95(mm)	Dice(%)	HD95(mm)	
nnU-Net	96.0 ± 0.9	3.7 ± 5.1	89.5 ± 1.9	4.9 ± 7.2	15h05m
SamMed2D zeroshot	91.7 ± 3.4	5.0 ± 1.7	74.4 ± 8.8	7.3 ± 2.4	0
SamMed2D finetune	92.7 ± 2.7	4.2 ± 1.6	79.5 ± 3.8	6.2 ± 2.0	2h47m
MedSAM zeroshot	91.8 ± 5.1	4.2 ± 1.5	24.8 ± 18.9	12.3 ± 2.8	0
MedSAM finetune	**97.2 ± 1.0**	**2.1 ± 0.5**	**91.9 ± 1.7**	**3.2 ± 0.9**	1h50m

matched conditions, but its performance dropped below 80% when tested with a 6 mm BB shift. Our findings also suggest that introducing more than 3 mm of BB variation at inference does not yield improved Dice performance compared to the nnU-Net baseline. However, MedSAM maintains a slight advantage in terms of HD when evaluated with a 6 mm BB perturbation, but only when it was fine-tuned under the same variation conditions. Based on these results, we chose to train and evaluate the proposed mcMedSAM model in the next section using BB perturbed with a 1 mm shift.

3.4 Multi-class MedSAM with single bounding box initialization

In this section, we evaluate the performance of the proposed mcMedSAM architecture, detailed in Section 2, which is designed to reduce user interaction to a single prompt for the simultaneous segmentation of the LV and MYO structures. To this end, we conducted an ablation study to assess the contribution of several key components of the architecture. The results are presented in Table 2. The first row of this table serves as the baseline and corresponds to a classical MedSAM model fine-tuned for the segmentation of the MYO only. The corresponding LV mask was derived through a post-processing pipeline based on anatomical priors, which infers the central region of the myocardium as the LV. This approach performs worse than the method reported in the last row of Table 1, not only because of the variability introduced by the bounding box prompt, but also because a single bounding box is used for both structures instead of one per structure. Although this strategy reduces the required user interaction, it also limits the anatomical priors available to guide segmentation, leading to less accurate delineation of the LV and MYO. Accordingly, we refined this approach



(a) Mean Dice score. (b) Mean HD.

Fig. 3: Impact of BB variation during training and inference on segmentation performance. Results are shown for a single fold of the cross-validation. nnU-Net results are included as a reference (dashed line).

to maintain high segmentation accuracy while relying on a single bounding box. Our ablation study first demonstrate the effectiveness of the structure token, which consistently improves segmentation performance by clarifying the target structure. The introduction of the decoupled architecture further enhanced segmentation accuracy, particularly for the MYO structure. Our final method mcMedSAM also improved the accuracy of both the LV and the MYO, obtaining our best results when using only one single BB. Overall, while the best configuration of our proposed model significantly outperforms conventional MedSAM fine-tuning in the single-prompt setting, its final segmentation performance remains comparable to that of the nnU-Net baseline. In terms of computational efficiency, our final method requires only 6 h 37 min of training, compared to 15 h 05 min for nnU-Net. However, at inference time, it takes about 24 seconds per volume with our method, whereas nnU-Net completes the same task in just 2.5 seconds.

4 Discussion and conclusion

In this study, we analyzed the performance of MedSAM and SamMed2D under both zero-shot and efficient fine-tuning configurations. Although SAM-based models were originally designed for zero-shot segmentation, our results highlight their limitations in handling complex anatomical structures, such as the myocardium in LGE MRI. Among the two models, MedSAM showed promising fine-tuning capabilities, even outperforming the state-of-the art nnU-Net, whereas the fine-tuning of SamMed2D led to only modest improvements.

Motivated by these findings, we focused our efforts on fine-tuning strategies for MedSAM. Given the known sensitivity of SAM models to prompt inputs, we quantified the impact of BB variations on MedSAM's performance both during

Table 2: Ablation study of the proposed mcMedSAM using a single BB with ±1 mm positional variation during training and inference.

Structure token	Decoupled Decoder	mcMedSAM	LV		MYO		Training time
			Dice(%)	HD95(mm)	Dice(%)	HD95(mm)	
–	–	–	90.8 ± 4.1	5.2 ± 2.5	86.5 ± 2.7	5.2 ± 2.0	1h49m
✓	–	–	94.2 ± 3.1	**3.6 ± 1.5**	88.2 ± 2.2	4.6 ± 1.7	1h50m
✓	–	✓	93.8 ± 3.2	3.8 ± 1.6	88.7 ± 2.3	4.3 ± 1.5	2h52m
✓	✓	–	94.1 ± 4.7	3.9 ± 2.8	88.5 ± 4.4	4.5 ± 2.0	4h46m
✓	✓	✓	**94.4 ± 3.7**	3.6 ± 1.9	**89.4 ± 3.0**	**4.0 ± 1.8**	6h37m

training and inference. Our empirical analysis revealed the critical importance of aligning BB perturbations introduced during fine-tuning with those expected at inference time. We also observed a sharp decline in segmentation accuracy as prompt quality deteriorated, with a critical threshold around 3 mm beyond which performance dropped below that of nnU-Net.

In addition, the need for precise BB initialization for each target structure remains a significant limitation, particularly in the context of clinical deployment, where such interactions are time-consuming and impractical. To address this issue, we proposed mcMedSAM, an extension of the SAM architecture specifically designed for multi-class segmentation from a single BB. While mcMedSAM improves segmentation in the single-prompt setting through standard MedSAM fine-tuning, its performance still converges toward that of nnU-Net and remains dependent on accurate prompt initialization.

Based on these results, we conclude that medical SAM models offer limited benefit for the segmentation of cardiac structures in LGE MRI. Although we successfully developed a method enabling multi-class segmentation from a single bounding box, the required positioning accuracy of the prompt does not provide a clear advantage over conventional approaches such as nnU-Net.

Acknowledgments. This work was supported by the French National Research Agency (PEPR digital Health).

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

References

1. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=plKu2GByCNW
2. Cheng, D., Qin, Z., Jiang, Z., Zhang, S., Lao, Q., Li, K.: Sam on medical images: A comprehensive study on three prompt modes. arXiv preprint arXiv:2305.00035 (2023)
3. 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)
4. Gao, Y., Xia, W., Hu, D., Wang, W., Gao, X.: Desam: Decoupled segment anything model for generalizable medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 509–519. Springer (2024)
5. Goujat, C.: A three-stage nnu-net framework for automated myocardial infarction quantification in lge-mri (2024), https://humanheart-project.creatis.insa-lyon.fr/myosaiqResources/2024_proceedings_myosaiq_celia.pdf, proceedings of the MyoSAIQ Workshop, 2024
6. 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. arXiv preprint arXiv:2304.09324 (2023)
7. Huang, P., Hu, S., Peng, B., Gong, X., Yin, P., Zhu, H., Wu, X., Wang, X.: Diffusion-empowered autoprompt medsam. arXiv preprint arXiv:2502.06817 (2025)
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. 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. In: IEEE International Conference on Computer Vision (ICCV). pp. 4015–4026 (2023)
10. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (Jan 2024). https://doi.org/10.1038/s41467-024-44824-z, https://doi.org/10.1038/s41467-024-44824-z
11. 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. IEEE (2023)
12. Roy, S., Wald, T., Koehler, G., Rokuss, M.R., Disch, N., Holzschuh, J., Zimmerer, D., Maier-Hein, K.H.: Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. arXiv preprint arXiv:2304.05396 (2023)
13. Stein, J., Di Folco, M., Schnabel, J.A.: Influence of prompting strategies on segment anything model (sam) for short-axis cardiac mri segmentation. In: BVM Workshop. pp. 54–59. Springer (2024)
14. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in neural information processing systems **33**, 7537–7547 (2020)
15. Xie, B., Tang, H., Duan, B., Cai, D., Yan, Y.: Masksam: Towards auto-prompt sam with mask classification for medical image segmentation. arXiv preprint arXiv:2403.14103 (2024)