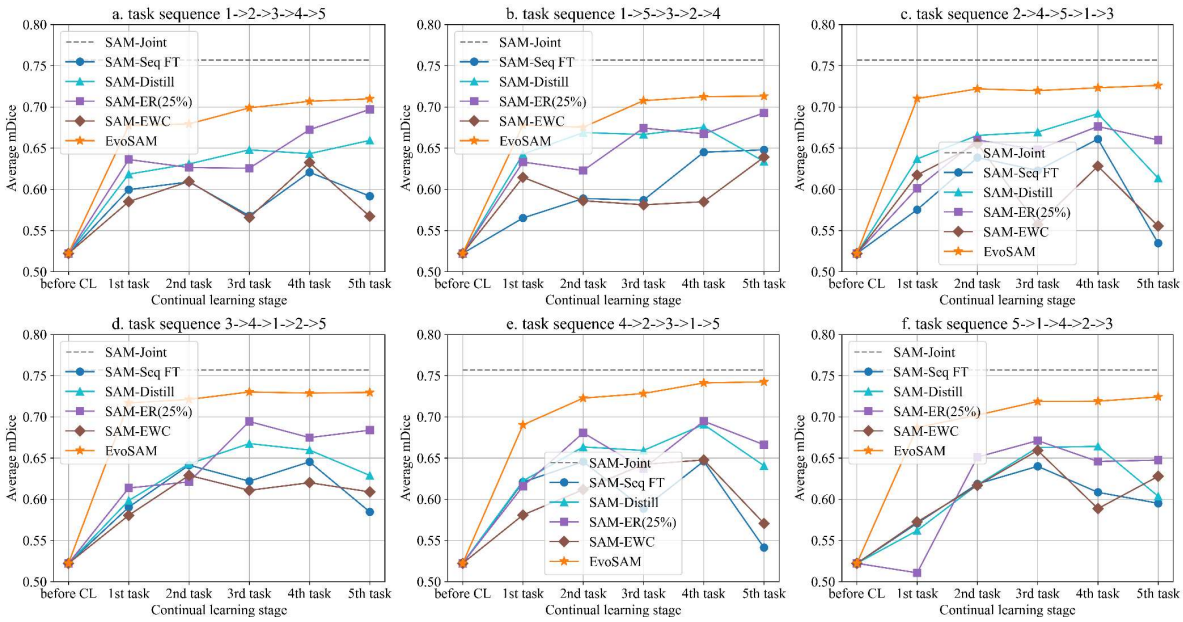
Supplementary Information

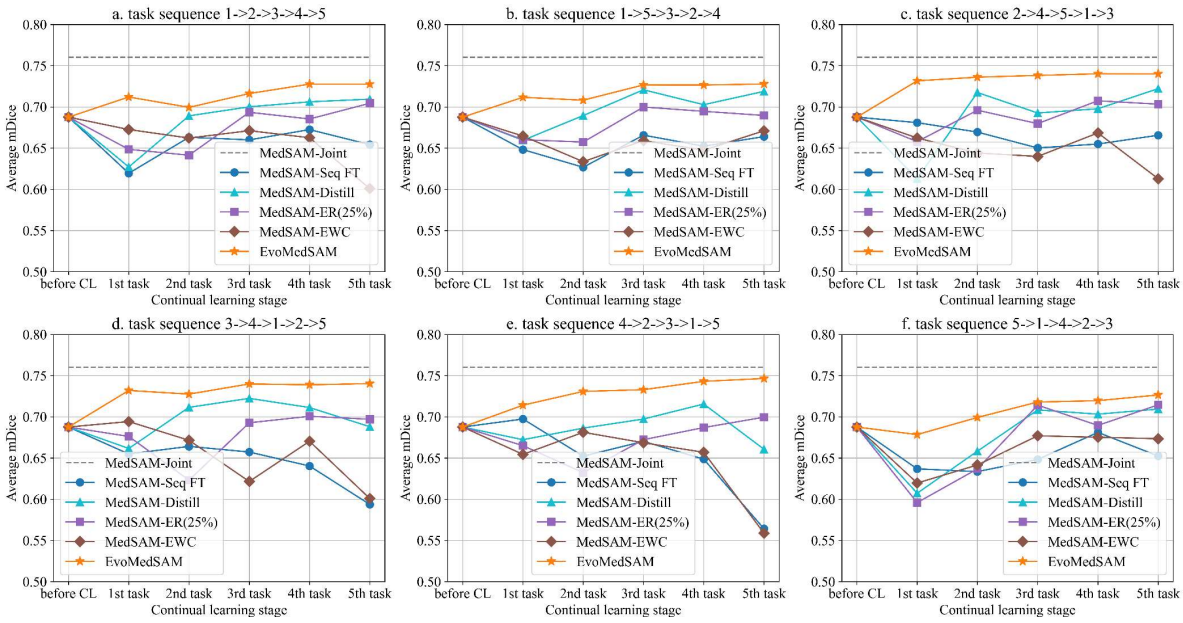
Supplementary Results

Average mDice Curve During Continual Learning Process

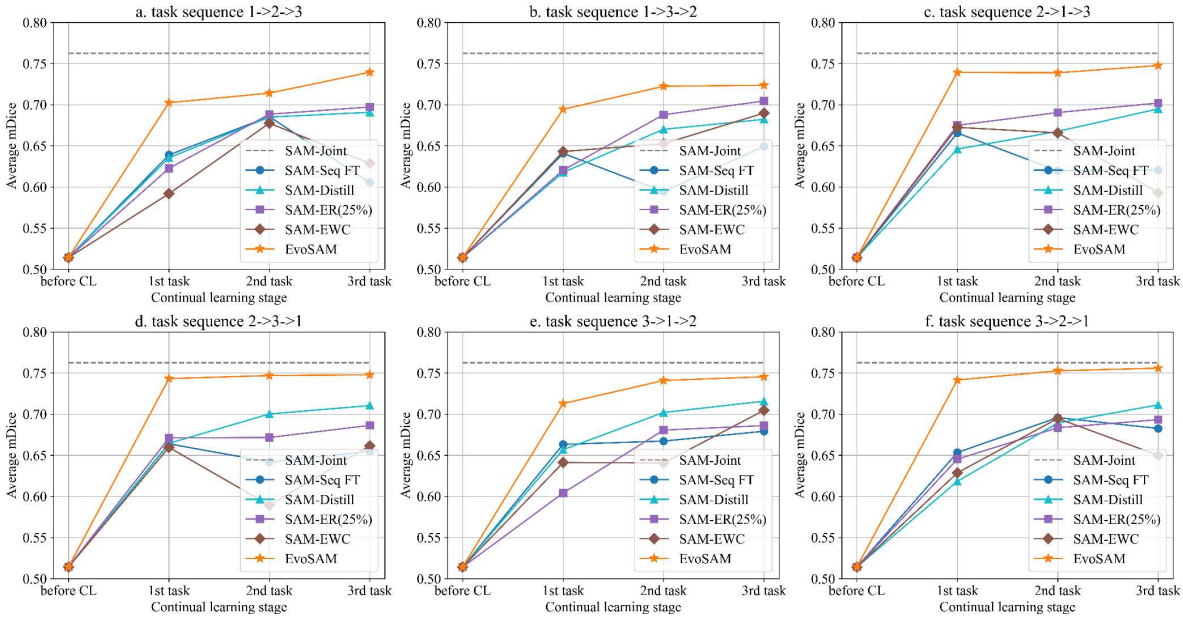
In this section, we report the mDice curves throughout the continual learning process for our method and baseline approaches (Seq FT, Distill[1], ER[2], and EWC[3]) across all experimental settings considered in this study (as shown in Supplementary Figures 1–5). As highlighted in the Results section, the baseline methods exhibit fluctuations in many cases under various experimental conditions, whereas our method consistently demonstrates an almost monotonic increasing trend. Moreover, in most cases, it achieves superior final performance compared to the baselines.



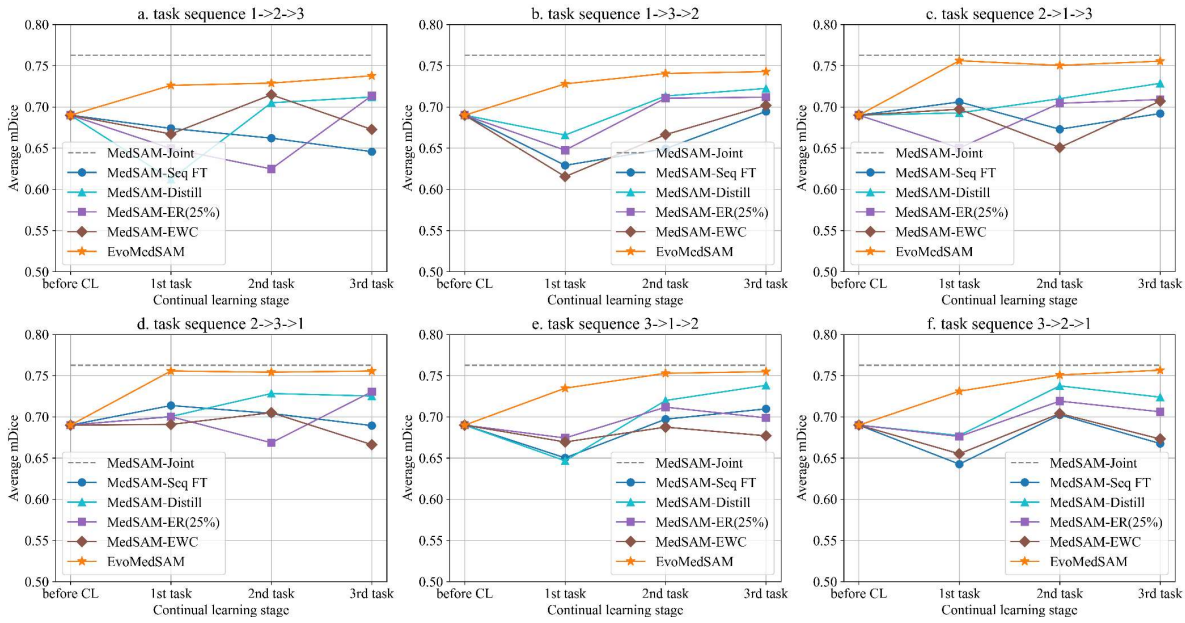
Supplementary Figure 1: Average mDice of blood vessel segmentation by our method and baseline methods for continual learning with SAM[4] under six different task sequences of the 5-task protocol.



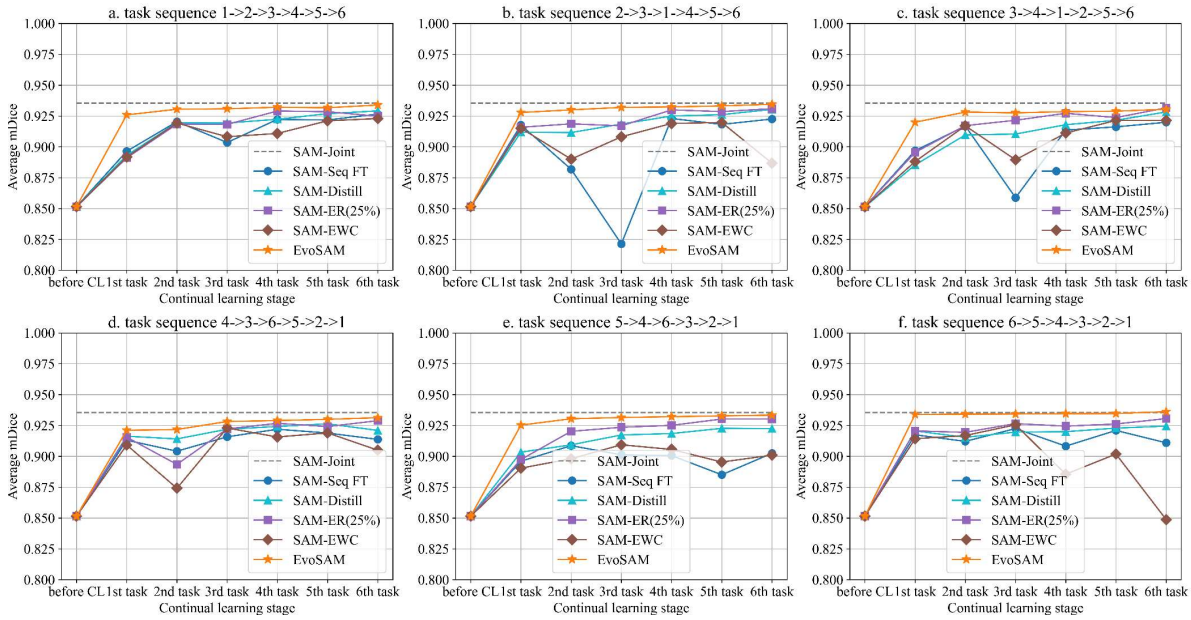
Supplementary Figure 2: Average mDice of blood vessel segmentation by our method and baseline methods for continual learning with MedSAM[5] under six different task sequences of the 5-task protocol.



Supplementary Figure 3: Average mDice of blood vessel segmentation by our method and baseline methods for continual learning with SAM under six different task sequences of the 3-task protocol.



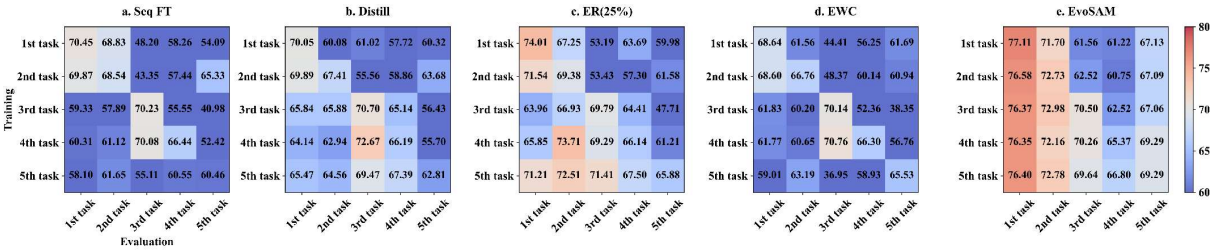
Supplementary Figure 4: Average mDice of blood vessel segmentation by our method and baseline methods for continual learning with MedSAM under six different task sequences of the 3-task protocol.



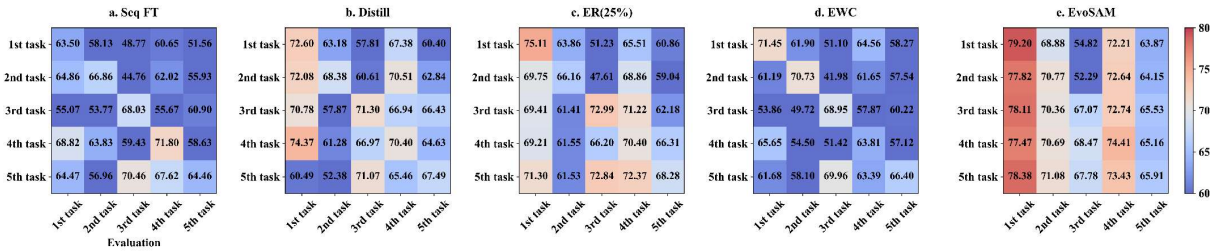
Supplementary Figure 5: Average mDice of multi-site prostate MRI segmentation by our method and baseline methods for continual learning with SAM under six different task sequences.

mDice Matrix During Continual Learning Process

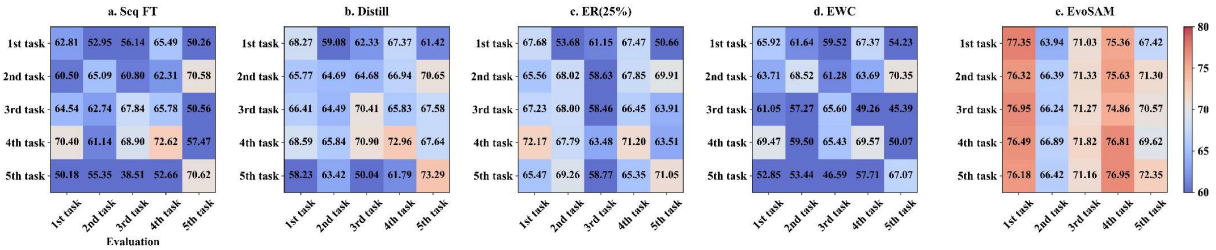
In this section, we report the mDice matrices for our method and baseline approaches (Seq FT, Distill, ER, and EWC) across all experimental settings (as shown in Supplementary Figures 6–35). The value in the i-th row and j-th column of the matrix represents the mDice score of the model on the j-th task's test set after continual learning with data from the 1st to the i-th task. As mentioned in the Results section, the columns of the lower triangular matrix for the baseline methods often exhibit a decreasing trend (or significant fluctuations), a typical manifestation of forgetting. In contrast, for our method, the values in the columns of the lower triangular matrix remain stable in the vast majority of experimental settings, and in some cases, even show an increasing trend (indicating improvement in performance on earlier tasks as new task knowledge is learned).



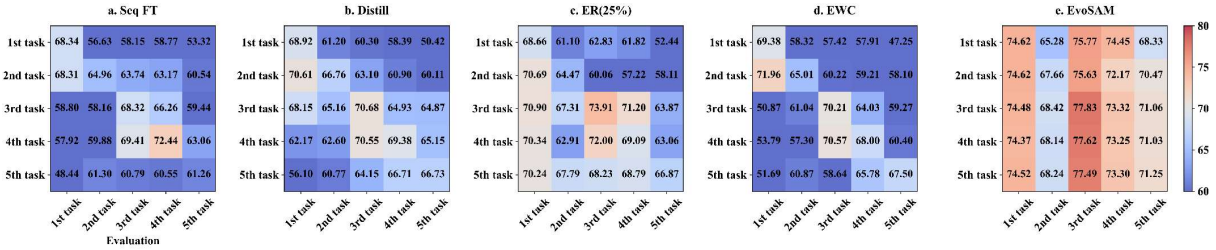
Supplementary Figure 6: The mDice matrices of EvoSAM and baseline methods are shown under the 5-task blood vessel segmentation protocol, following the task order: 1→2→3→4→5, labeled as the 1st to 5th tasks, respectively.



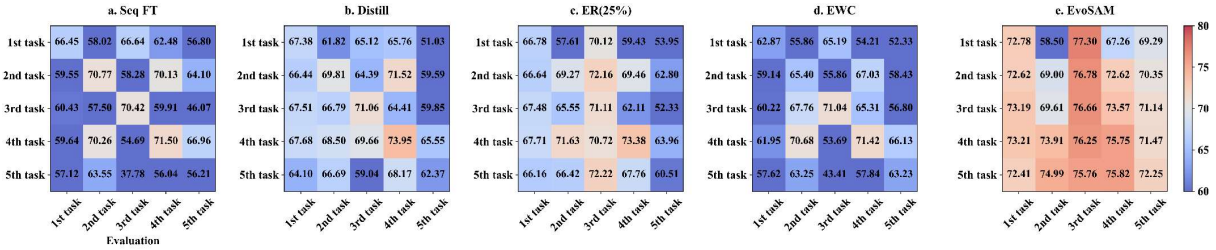
Supplementary Figure 7: The mDice matrices of EvoSAM and baseline methods are shown under the 5-task blood vessel segmentation protocol, following the task order: 1→5→3→2→4, labeled as the 1st to 5th tasks, respectively.



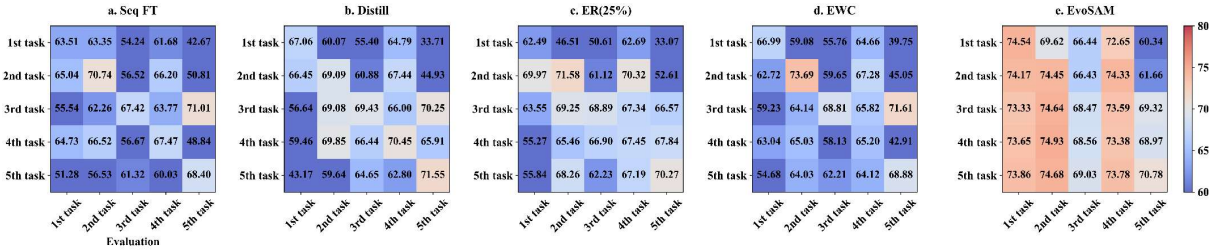
Supplementary Figure 8: The mDice matrices of EvoSAM and baseline methods are shown under the 5-task blood vessel segmentation protocol, following the task order: 2→4→5→1→3, labeled as the 1st to 5th tasks, respectively.



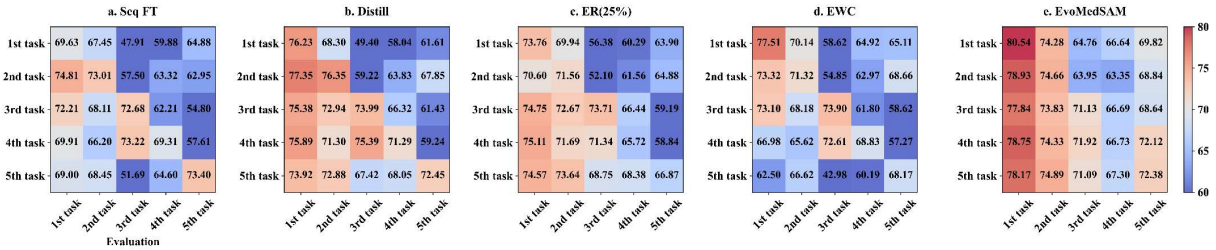
Supplementary Figure 9: The mDice matrices of EvoSAM and baseline methods are shown under the 5-task blood vessel segmentation protocol, following the task order: 3→4→1→2→5, labeled as the 1st to 5th tasks, respectively.



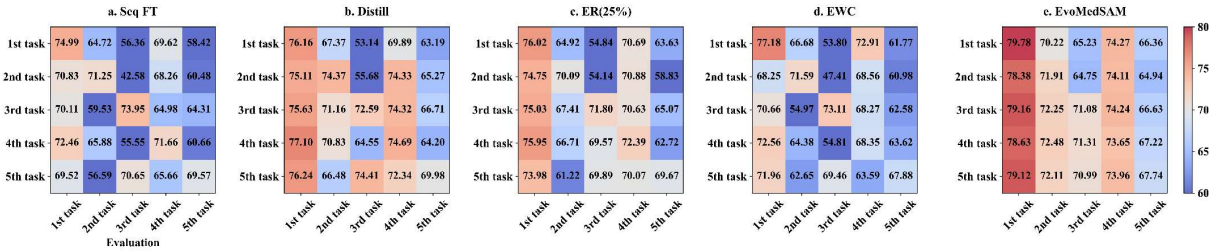
Supplementary Figure 10: The mDice matrices of EvoSAM and baseline methods are shown under the 5-task blood vessel segmentation protocol, following the task order: 4→2→3→1→5, labeled as the 1st to 5th tasks, respectively.



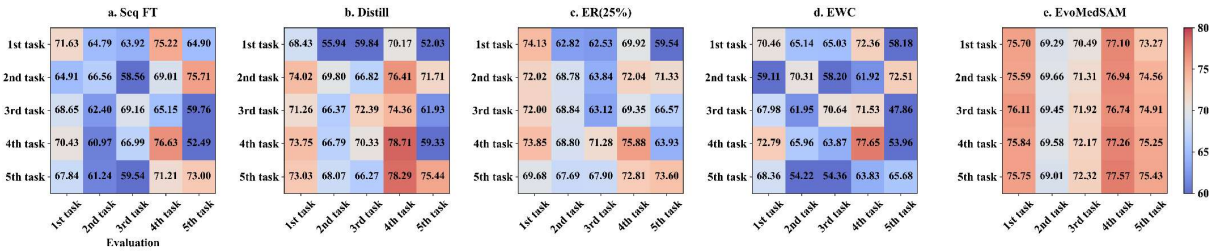
Supplementary Figure 11: The mDice matrices of EvoSAM and baseline methods are shown under the 5-task blood vessel segmentation protocol, following the task order: 5→1→4→2→3, labeled as the 1st to 5th tasks, respectively.



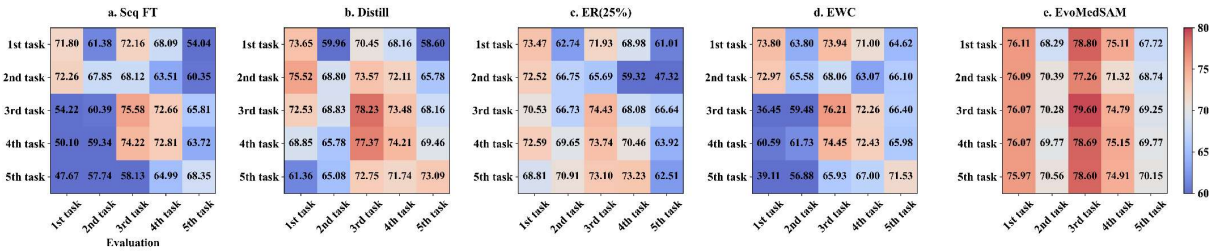
Supplementary Figure 12: The mDice matrices of EvoMedSAM and baseline methods are shown under the 5-task blood vessel segmentation protocol, following the task order: 1→2→3→4→5, labeled as the 1st to 5th tasks, respectively.



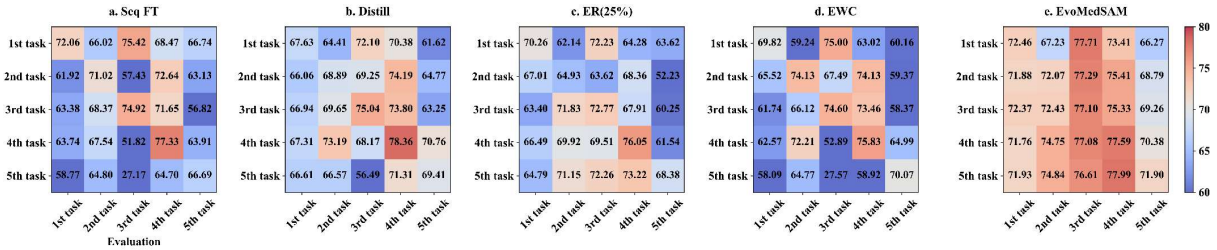
Supplementary Figure 13: The mDice matrices of EvoMedSAM and baseline methods are shown under the 5-task blood vessel segmentation protocol, following the task order: 1→5→3→2→4, labeled as the 1st to 5th tasks, respectively.



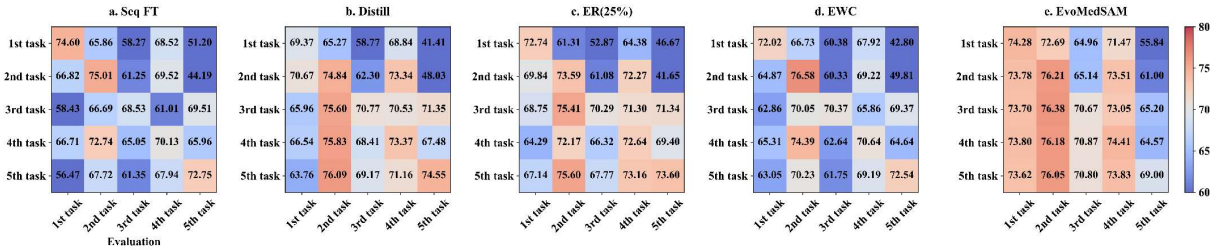
Supplementary Figure 14: The mDice matrices of EvoMedSAM and baseline methods are shown under the 5-task blood vessel segmentation protocol, following the task order: 2→4→5→1→3, labeled as the 1st to 5th tasks, respectively.



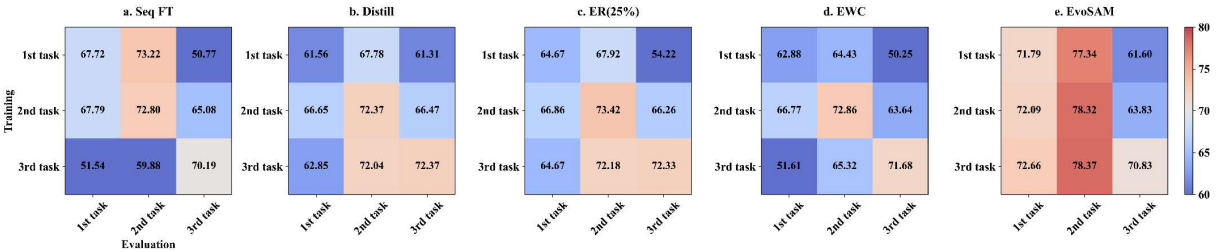
Supplementary Figure 15: The mDice matrices of EvoMedSAM and baseline methods are shown under the 5-task blood vessel segmentation protocol, following the task order: 3→4→1→2→5, labeled as the 1st to 5th tasks, respectively.



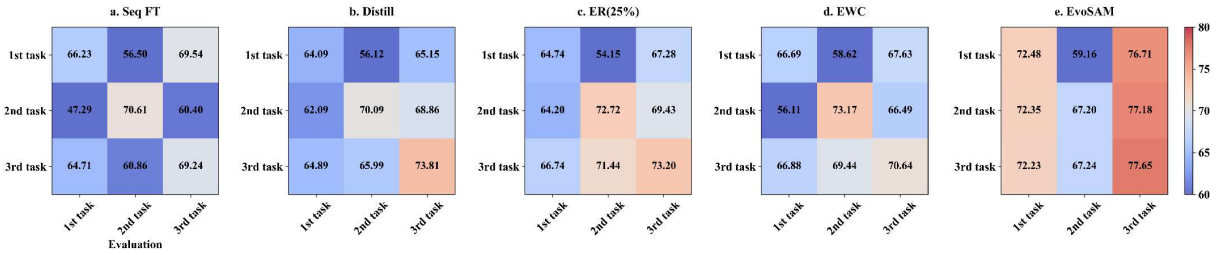
Supplementary Figure 16: The mDice matrices of EvoMedSAM and baseline methods are shown under the 5-task blood vessel segmentation protocol, following the task order: 4→2→3→1→5, labeled as the 1st to 5th tasks, respectively.



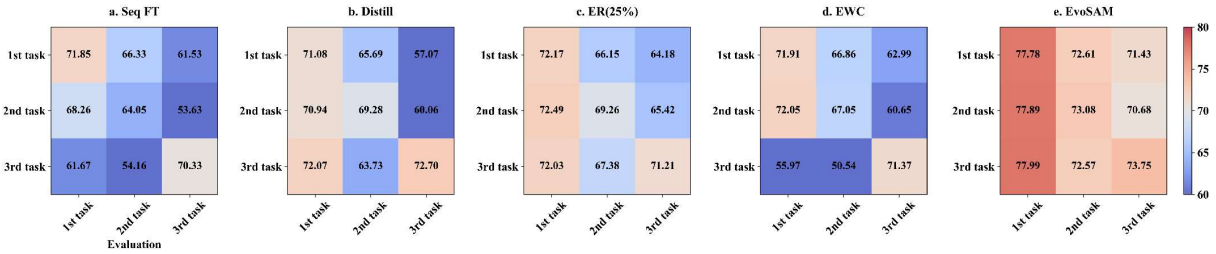
Supplementary Figure 17: The mDice matrices of EvoMedSAM and baseline methods are shown under the 5-task blood vessel segmentation protocol, following the task order: 5→1→4→2→3, labeled as the 1st to 5th tasks, respectively.



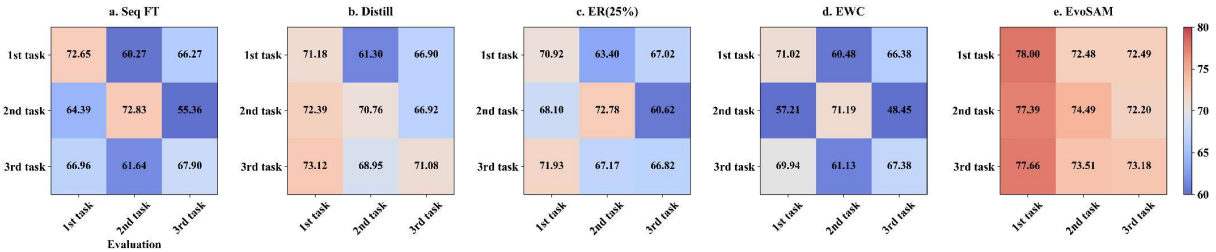
Supplementary Figure 18: The mDice matrices of EvoSAM and baseline methods are shown under the 3-task blood vessel segmentation protocol, following the task order: 1→2→3, labeled as the 1st to 3rd tasks, respectively.



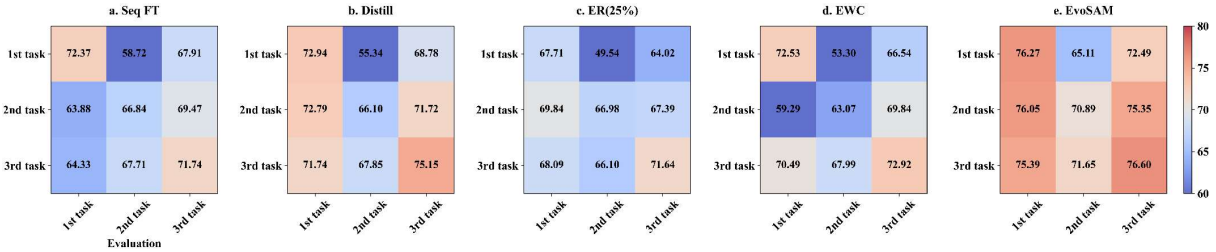
Supplementary Figure 19: The mDice matrices of EvoSAM and baseline methods are shown under the 3-task blood vessel segmentation protocol, following the task order: 1→3→2, labeled as the 1st to 3rd tasks, respectively.



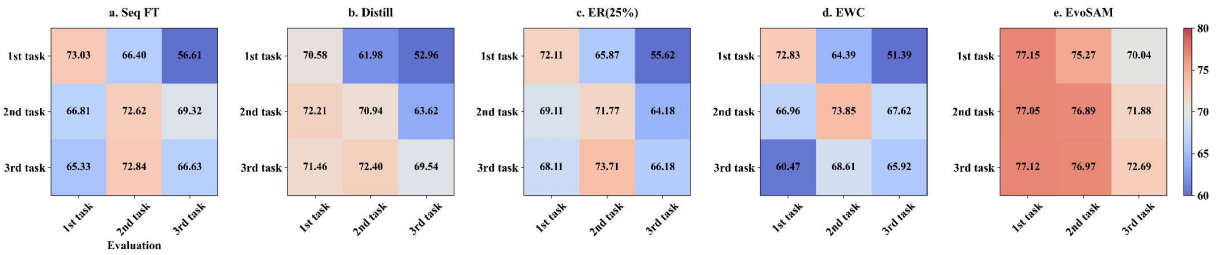
Supplementary Figure 20: The mDice matrices of EvoSAM and baseline methods are shown under the 3-task blood vessel segmentation protocol, following the task order: 2→1→3, labeled as the 1st to 3rd tasks, respectively.



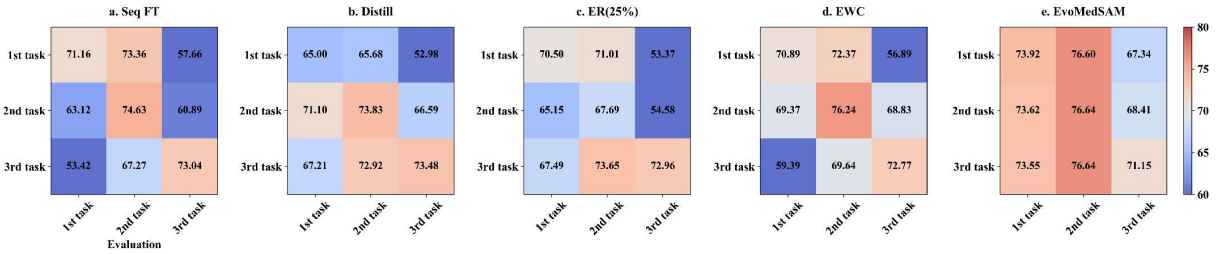
Supplementary Figure 21: The mDice matrices for different continual learning models based on SAM are shown under the 3-task blood vessel segmentation protocol, following the task order: 2→3→1, labeled as the 1st to 3rd tasks, respectively.



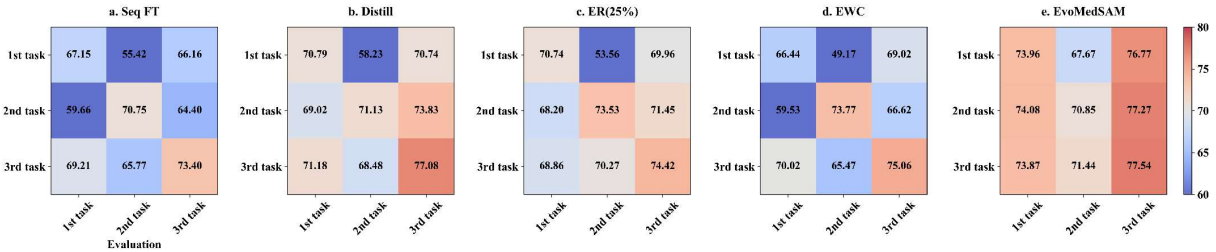
Supplementary Figure 22: The mDice matrices of EvoSAM and baseline methods are shown under the 3-task blood vessel segmentation protocol, following the task order: 3→1→2, labeled as the 1st to 3rd tasks, respectively.



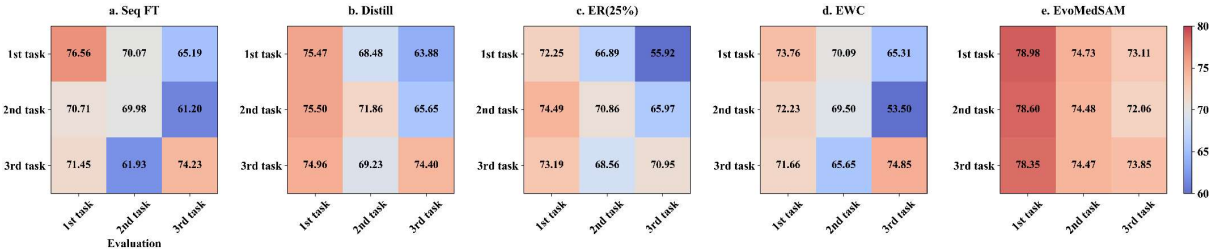
Supplementary Figure 23: The mDice matrices of EvoSAM and baseline methods are shown under the 3-task blood vessel segmentation protocol, following the task order: 3→2→1, labeled as the 1st to 5th tasks, respectively.



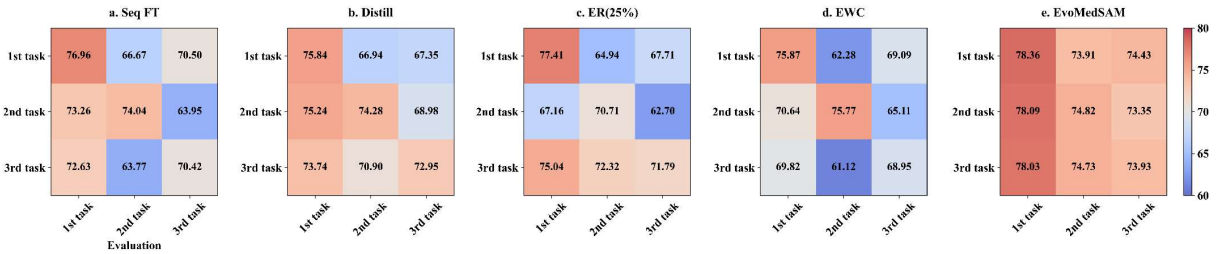
Supplementary Figure 24: The mDice matrices of EvoMedSAM and baseline methods are shown under the 3-task blood vessel segmentation protocol, following the task order:1→2→3, labeled as the 1st to 3rd tasks, respectively.



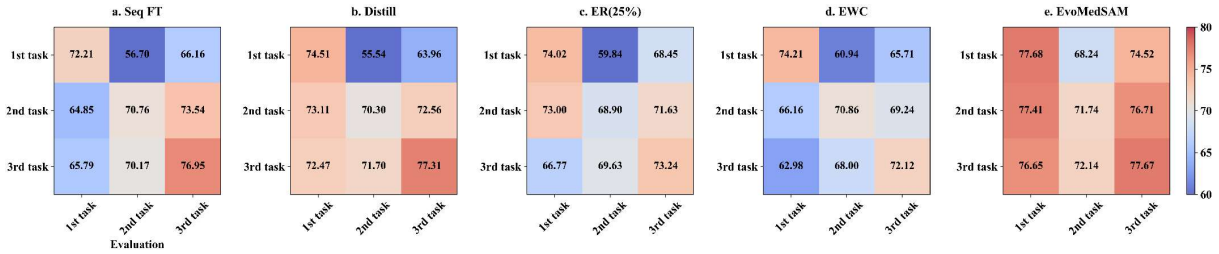
Supplementary Figure 25: The mDice matrices of EvoMedSAM and baseline methods are shown under the 3-task blood vessel segmentation protocol, following the task order: 1→3→2, labeled as the 1st to 3rd tasks, respectively.



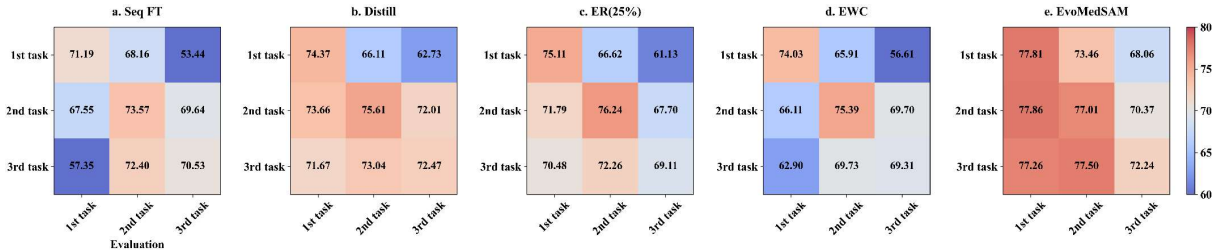
Supplementary Figure 26: The mDice matrices of EvoMedSAM and baseline methods are shown under the 3-task blood vessel segmentation protocol, following the task order: 2→1→3, labeled as the 1st to 3rd tasks, respectively.



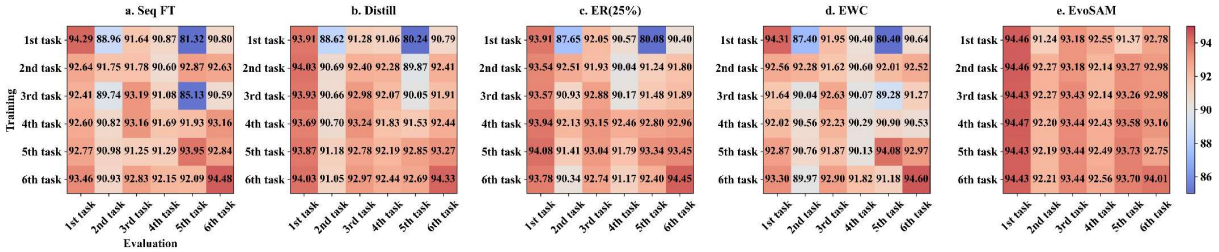
Supplementary Figure 27: The mDice matrices of EvoMedSAM and baseline methods are shown under the 3-task blood vessel segmentation protocol, following the task order: 2→3→1, labeled as the 1st to 3rd tasks, respectively.



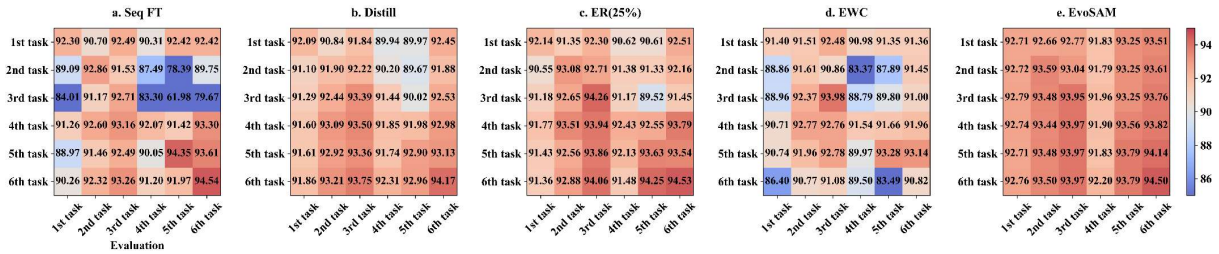
Supplementary Figure 28: The mDice matrices of EvoMedSAM and baseline methods are shown under the 3-task blood vessel segmentation protocol, following the task order: 3→1→2, labeled as the 1st to 3rd tasks, respectively.



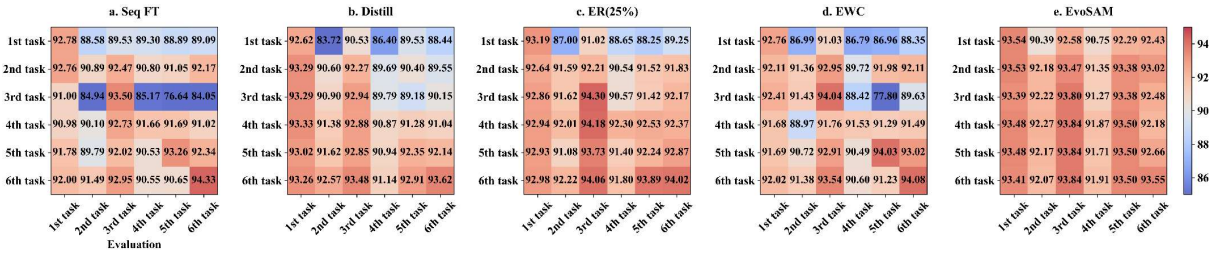
Supplementary Figure 29: The mDice matrices of EvoMedSAM and baseline methods are shown under the 3-task blood vessel segmentation protocol, following the task order: 3→2→1, labeled as the 1st to 3rd tasks, respectively.



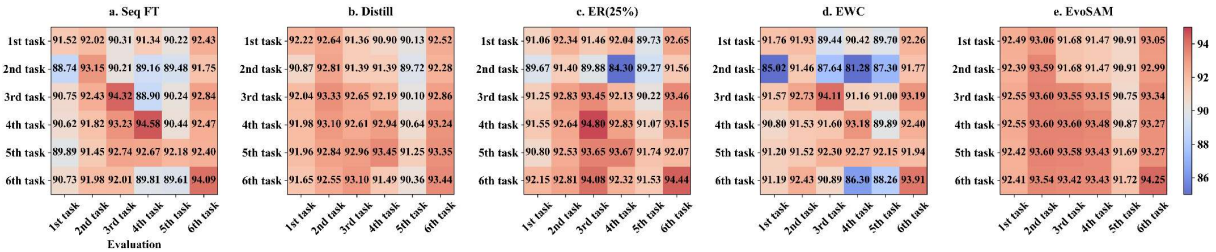
Supplementary Figure 30: The mDice matrices of EvoSAM and baseline methods are shown under the multi-site prostate MRI segmentation, following the task order: 1→2→3→4→5→6, labeled as the 1st to 6th tasks, respectively.



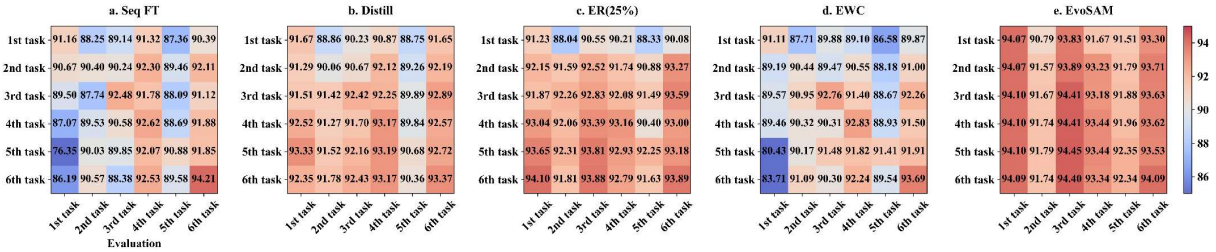
Supplementary Figure 31: The mDice matrices of EvoSAM and baseline methods are shown under the multi-site prostate MRI segmentation, following the task order: 2→3→1→4→5→6, labeled as the 1st to 6th tasks, respectively.



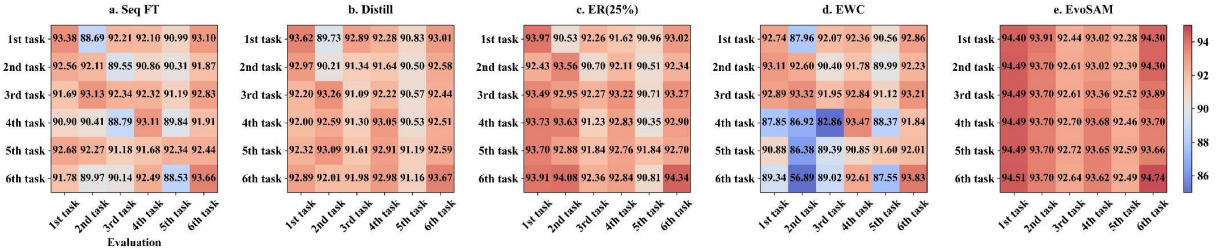
Supplementary Figure 32: The mDice matrices of EvoSAM and baseline methods are shown under the multi-site prostate MRI segmentation, following the task order: 3→4→1→2→5→6, labeled as the 1st to 6th tasks, respectively.



Supplementary Figure 33: The mDice matrices of EvoSAM and baseline methods are shown under the multi-site prostate MRI segmentation, following the task order: 4→3→6→5→2→1, labeled as the 1st to 6th tasks, respectively.



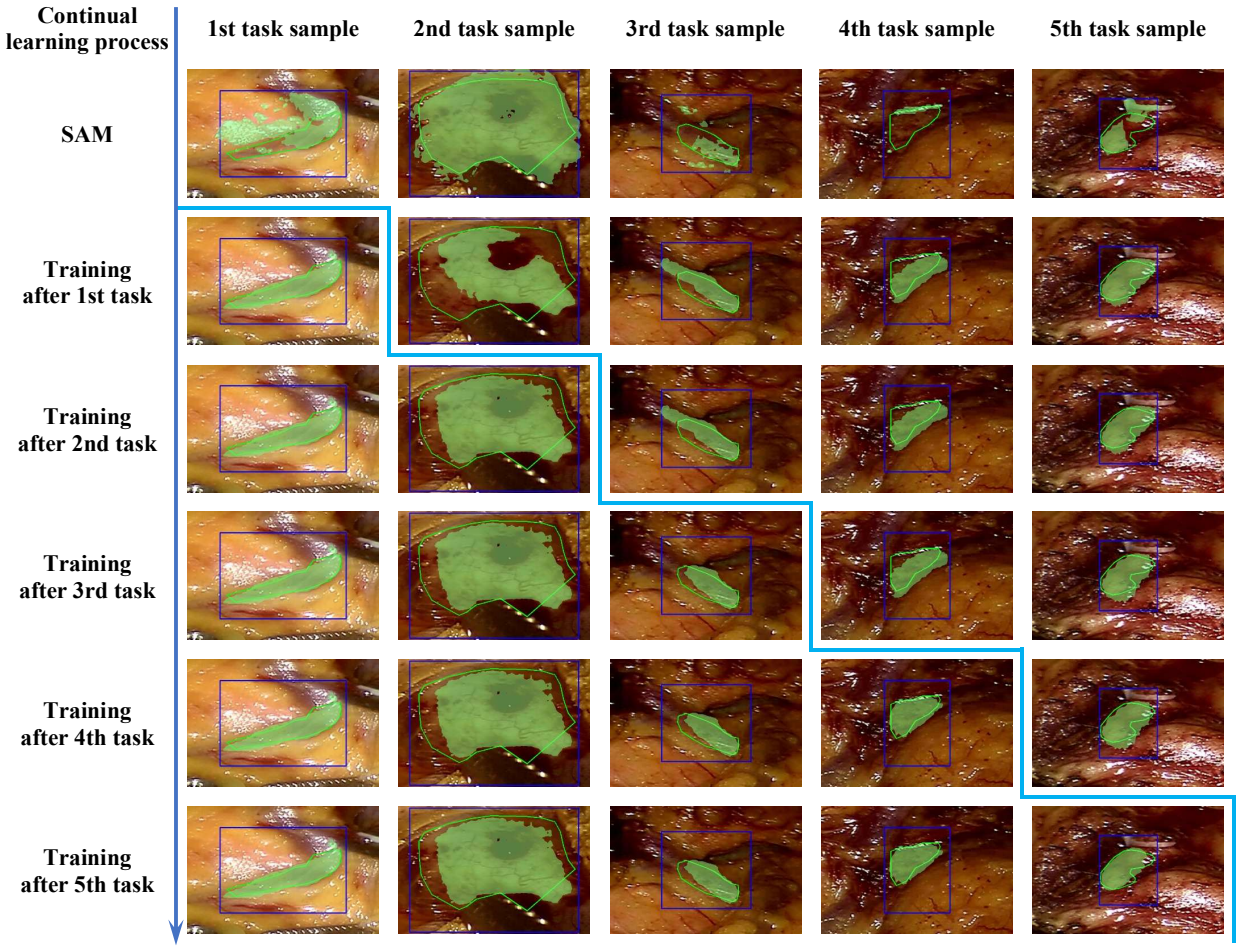
Supplementary Figure 34: The mDice matrices of EvoSAM and baseline methods are shown under the multi-site prostate MRI segmentation, following the task order: 5→4→6→3→2→1, labeled as the 1st to 6th tasks, respectively.



Supplementary Figure 35: The mDice matrices of EvoSAM and baseline methods are shown under the multi-site prostate MRI segmentation, following the task order: 6→5→4→3→2→1, labeled as the 1st to 6th tasks, respectively.

Test Cases

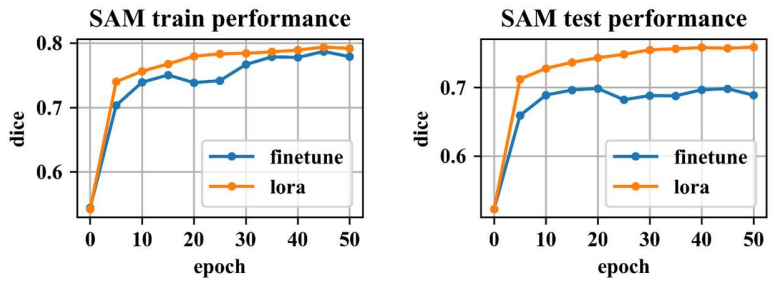
To better illustrate the effectiveness of EvoSAM during continual learning, we present its segmentation results for five surgical images across five tasks with a task sequence of 4→2→3→1→5 under the 5-task protocol. As illustrated in Supplementary Figures 44, before continual learning, SAM exhibits relatively satisfactory results on the image of the 2nd task, though with irregular and uneven boundaries. Conversely, its segmentation results for the images of the 1st, 3rd, and 5th tasks are dispersed and imprecise, with only a minor portion of the blood vessels being accurately segmented in the 4th task. After learning on the 1st task, EvoSAM demonstrates significantly enhanced segmentation precision for the corresponding surgical images. We also observe noticeable improvements in the segmentation of surgical images from unseen tasks, specifically the 3rd, 4th, and 5th tasks. An exception arises with the 2nd task, where the segmented regions show significant shrinkage, and a substantial portion of the blood vessel area is erroneously classified as background. However, after learning on the 2nd task, the segmentation performance on its image improves considerably, capturing more blood vessel areas with relatively smooth boundaries, while maintaining robust and accurate segmentation results for the 1st task. This trend continues consistently throughout the subsequent phases of continual learning. These results further demonstrate the capability of our EvoSAM system to effectively accumulate new knowledge while minimizing the forgetting of previously acquired knowledge.



Supplementary Figure 44: Visualization of blood vessel segmentation by EvoSAM after learning each task during continual learning. The segmentation results after EvoSAM learns each task are shown for five surgical images across five tasks with a task sequence of 4(1st)→2(2nd)→3(3rd)→1(4th)→5(5th)) under the 5-task protocol.

EvoSAM Mitigates Progressive Overfitting

Due to the large number of parameters in the SAM model, fine-tuning often leads to overfitting. In our experiments on blood vessel segmentation, we observed that LoRA fine-tuning of SAM achieves performance on the training set comparable to full-parameter fine-tuning (the image encoder and mask encoder) while yielding superior performance on the test set, as shown in Supplementary Figure 36, this suggests that, to some extent, LoRA fine-tuning can mitigate the overfitting issue of SAM in this task.

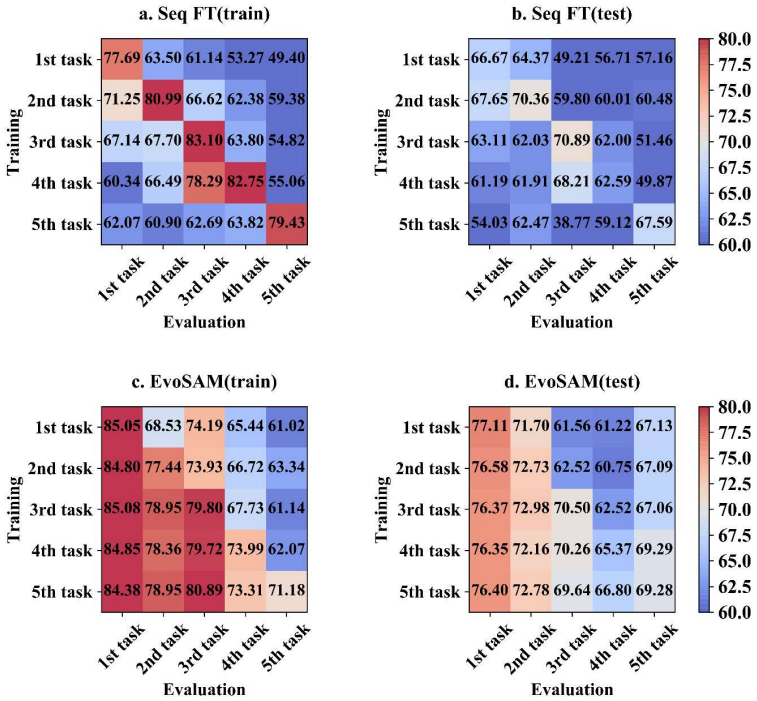


Supplementary Figure 36: LoRA alleviates the overfitting issue of SAM fine-tuning in the blood vessel segmentation.

In the continual learning setting, sequential fine-tuning of SAM often results in more severe progressive overfitting[6]. Supplementary Figures 36a and 36b present the mDice matrices of SAM sequential fine-tuning on the training and test sets, respectively, under the 5-task blood vessel segmentation protocol, following the task order: 1→2→3→4→5. Notably, the diagonal elements of the training set mDice matrix are significantly higher than those in other regions and markedly exceed the corresponding diagonal elements in the test set mDice matrix. This indicates that the SAM model overfits sequentially to each task's training set, resulting in suboptimal performance not only on past and future tasks but even on the most recently learned task.

In contrast, EvoSAM employs a LoRA-based localized parameter isolation strategy to preserve the generalization ability of the pre-trained model, substantially mitigating this overfitting. As shown in Supplementary Figures 37c and 37d, the gap between the diagonal elements of the training and test set mDice matrices is significantly reduced. Consequently, despite adjusting fewer parameters and sometimes

achieving a lower fit on the training set of the most recently learned task, EvoSAM often achieves superior test performance by effectively integrating prior knowledge from the pre-trained model with information from the newly learned task.



Supplementary Figure 37: The mDice matrices in training set and testing set of EvoSAM and Seq FT based on SAM are shown under the 5-task blood vessel segmentation protocol, following the task order: 1→2→3→4→5, labeled as the 1st to 5th tasks, respectively.

Ablation Study

We conducted ablation studies on EvoSAM and EvoMedSAM based on four key components under the 5-task and 3-task blood vessel segmentation protocols: (a) LoRA: fine-tuning the image encoder and mask decoder of the SAM model using Low-Rank Adaptation[7] (by default, full-parameter fine-tuning is applied to both the image encoder and mask decoder); (b) First Task Adaptation: fine-tuning the image encoder only on the first task, while keeping the image encoder fully frozen for subsequent tasks; (c) Parameter Isolation: independently learning task-specific mask decoder expert parameters for different tasks; and (d)

Expert Matching: applying an expert matching strategy after parameter isolation. Specifically, the nearest neighbor (NN) algorithm was used as a baseline for matching, while Ridge Regression[8] was ingeniously applied in this study. Ground Truth refers to directly using true task labels for matching, which, while impractical in real-world scenarios due to the unavailability of task labels, provides an upper bound for the performance of the expert matching algorithms.

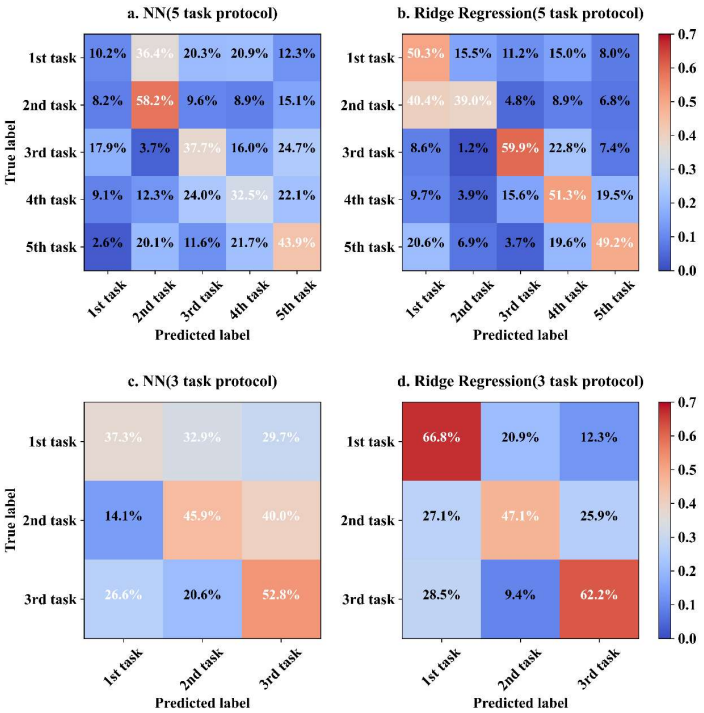
The experimental results, as shown in Supplementary Tables 1 and 2, demonstrate that compared to the default sequential fine-tuning setup, LoRA fine-tuning effectively improves the average mDice score and mitigates forgetting. When the first task adaptation strategy is further applied to the image encoder, parameter updates are restricted, leading to a greater reduction in forgetting during the continual learning process and further improvements in the average mDice (except for a slight decrease in EvoSAM performance under the 3-task blood vessel segmentation protocol, likely due to reduced model plasticity). When parameter isolation is additionally applied to the mask decoder, forgetting during continual learning is further alleviated; however, the expert matching strategy becomes crucial for achieving optimal performance (average mDice). Using the nearest neighbor algorithm for matching results in suboptimal average mDice, while the ridge regression approach further improves the average mDice, approaching the performance achieved with ground truth labels.

Supplementary Table 1: Ablation Study for EvoSAM and EvoMedSAM under the 5-task blood vessel segmentation protocol. The average mDice and average forgetting over 6 task orders are reported.

Pretrained Model	LoRA	First Task Adaption	Parameter Isolation	Expert Matching	Average mDice	Average Forgetting
SAM					58.26±3.77	3.87±1.41
	√				69.67±2.42	2.91±0.87
	√	√			70.50±1.65	1.55±0.92
	√	√	√	NN	70.42±1.61	0.35±0.40
	√	√	√	Ridge Regression	**72.42±1.08**	**0.04±0.23**
	√	√	√	Ground Truth	73.46±0.65	0.00±0.00
MedSAM					63.24±3.90	4.20±1.90
	√				69.23±3.61	3.01±1.30
	√	√			73.20±0.72	1.00±0.52
	√	√	√	NN	73.06±1.27	0.31±0.26
	√	√	√	Ridge Regression	**73.49±0.78**	**0.14±0.19**
	√	√	√	Ground Truth	74.61±0.85	0.00±0.00

Supplementary Table 2: Ablation Study for EvoSAM and EvoMedSAM under the 3-task blood vessel segmentation protocol. The average mDice and average forgetting over 6 task orders are reported.

Pretrained Model	LoRA	First Task Adaptation	Parameter Isolation	Expert Matching	Average mDice	Average Forgetting
SAM					64.87±2.83	5.73±1.56
	√				**74.63±1.13**	1.58±0.66
	√	√			74.19±0.92	0.62±0.75
	√	√	√	NN	73.77±0.95	0.17±0.18
	√	√	√	Ridge Regression	74.34±1.00	**0.07±0.23**
	√	√	√	Ground Truth	74.83±0.76	0.00±0.00
MedSAM					68.32±2.08	4.75±1.76
	√				74.51±0.96	1.84±0.47
	√	√			74.53±0.78	0.72±0.29
	√	√	√	NN	74.18±0.78	0.22±0.20
	√	√	√	Ridge Regression	**75.06±0.74**	**0.11±0.14**
	√	√	√	Ground Truth	75.28±0.37	0.00±0.00

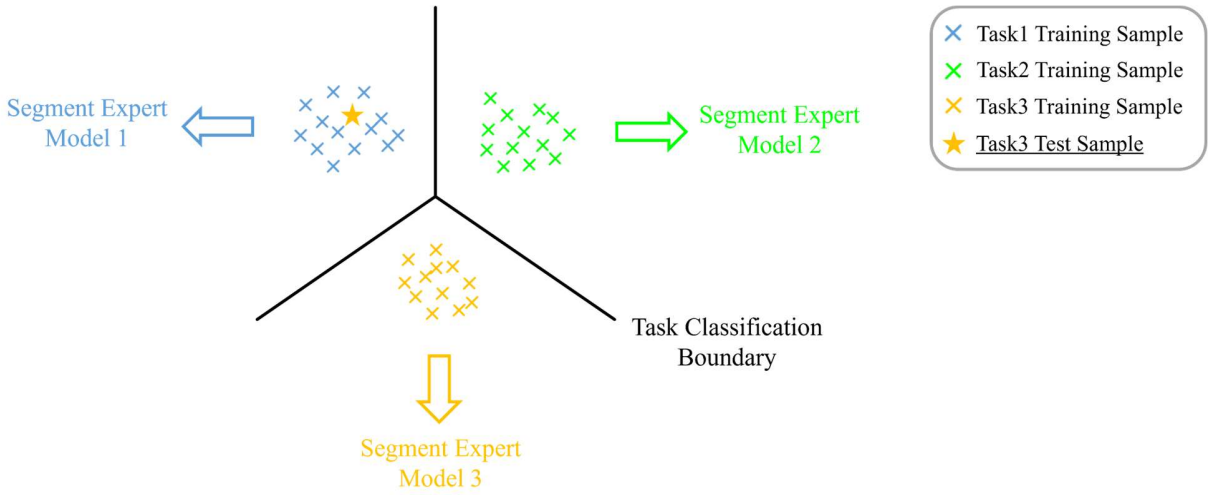


Supplementary Figure 38: The confusion matrix of NN and Ridge Regression for expert matching under 5-task and 3-task blood vessel segmentation protocol following the task order: 1→2→3→4→5 and 1→2→3.

We further visualized the confusion matrices of expert matching using NN and Ridge Regression under the 5-task and 3-task blood vessel segmentation protocols, following the default task order (1→2→3→4→5 and 1→2→3), as shown in Supplementary Figure 38. The results indicate that Ridge Regression achieves significantly higher matching accuracy compared to NN (except for the 2nd task in the 5-task protocol). This aligns with the conclusion that expert matching using Ridge Regression results in lower forgetting and higher final average mDice performance (as shown in Supplementary Tables 1 and 2).

However, the matching accuracy of Ridge Regression remains relatively low, despite achieving a final segmentation performance close to that obtained by directly using ground truth labels. We believe this is due to the fact that, even when Ridge Regression classifies samples into incorrect tasks, it may still select an appropriate model. As shown in Supplementary Figure 39, segment expert models 1, 2, and 3 are trained using the training samples of tasks 1, 2, and 3, respectively. Additionally, an expert matcher is trained on

these samples for task classification and expert matching, forming task classification boundaries in the sample feature space. For an outlier test sample from task 3, it may be closer to the training data distribution of task 1, leading to misclassification as task 1. However, segment expert model 1, trained on task 1's training data, will likely produce a good segmentation result for this sample.



Supplementary Figure 39: Outlier task identification and expert matching.

Supplementary Methods

Blood Vessel Segmentation Program

In this section, we detail the usage of our blood vessel segmentation program. The main interface of the program is shown in Supplementary Figure 40. For a given surgical image to be segmented, users can begin by selecting the target blood vessel region using a bounding box. The program then utilizes the model's prompt encoder and mask decoder to generate an initial segmentation result based on pre-computed image features (Supplementary Figure 41). Users can further refine the initial segmentation result by erasing extraneous regions using an adjustable eraser tool (Supplementary Figure 42) or adding missed regions

using an adjustable brush tool (Supplementary Figure 43). The eraser and brush tools can be used iteratively until the desired segmentation result is achieved. The total time spent on manual adjustments is automatically recorded for comparative analysis.

Supplementary Figure 40: Step 1, selection of the target blood vessel.

Supplementary Figure 41: Step 2, automated inference of the initial segmentation by the model.

Supplementary Figure 42: Step 3, refinement by erasing extraneous regions using a circular eraser.

Supplementary Figure 43: Step 4, refinement by adding missed regions using a circular brush.

References

1. Li, Z., & Hoiem, D. Learning without average forgetting. IEEE transactions on pattern analysis and machine intelligence, **40**, 2935-2947 (2017).

2. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P. K., Torr, P. H., & Ranzato, M. A. On tiny episodic memories in continual learning. arXiv preprint arXiv:1902.10486 (2019).
3. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R.. Overcoming catastrophic average forgetting in neural networks. Proceedings of the national academy of sciences, **114**, 3521-3526 (2017).
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 4015-4026 (2023).
5. Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. Segment anything in medical images. Nature Communications, **15**, 654 (2024).
6. Zhang, G., Wang, L., Kang, G., Chen, L., & Wei, Y. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 19148-19158) (2023).
7. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. Lora: Low-rank adaptation of large language models. Preprint at https://doi.org/10.48550/arXiv.2106.09685 (2021).
8. Murphy, K. P. Machine learning: a probabilistic perspective. MIT press (2012).
9. Liu, Q., Dou, Q., & Heng, P. A. Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23 (pp. 475-485). Springer International Publishing (2020).