Continual Alignment for SAM: Rethinking Foundation Models for Medical Image Segmentation in Continual Learning

Jiayi Wang[1]*, Wei Dai[1]*, Haoyu Wang[1]*, Sihan Yang[1]†, Haixia Bi[2]‡, Jian Sun[3],
[1]Xi'an Jiaotong University
[2]School of Information and Communications Engineering, Xi'an Jiaotong University
[3]School of Mathematics and Statistics, Xi'an Jiaotong University

Abstract

*In medical image segmentation, heterogeneous privacy policies across institutions often make joint training on pooled datasets infeasible, motivating continual image segmentation—learning from data streams without catastrophic forgetting. While the Segment Anything Model (SAM) offers strong zero-shot priors and has been widely fine-tuned across downstream tasks, its large parameter count and computational overhead challenge practical deployment. This paper demonstrates that the SAM paradigm is highly promising once its computational efficiency and performance can be balanced. To this end, we introduce the **Alignment Layer**, a lightweight, plug-and-play module which aligns encoder–decoder feature distributions to efficiently adapt SAM to specific medical images, improving accuracy while reducing computation. Building on SAM and the Alignment Layer, we then propose **C**ontinual **A**lignment for **SAM** (**CA-SAM**), a continual learning strategy that automatically adapts the appropriate Alignment Layer to mitigate catastrophic forgetting, while leveraging SAM's zero-shot priors to preserve strong performance on unseen medical datasets. Experimented across nine medical segmentation datasets under continual-learning scenario, CA-SAM achieves state-of-the-art performance. Our code, models and datasets will be released on https://github.com/azzzzyo/Continual-Alignment-for-SAM.*

1. Introduction

In medical image segmentation, privacy and governance constraints across institutions often preclude joint training on pooled datasets, making continual segmentation—learning from sequential data streams without revisiting prior samples—practically essential yet technically challenging. Meanwhile, the Segment Anything Model (SAM [17]) has demonstrated strong segmentation performance and *zero-shot* capability. Its potential in continual learning settings warrants further exploration [7, 35, 40].

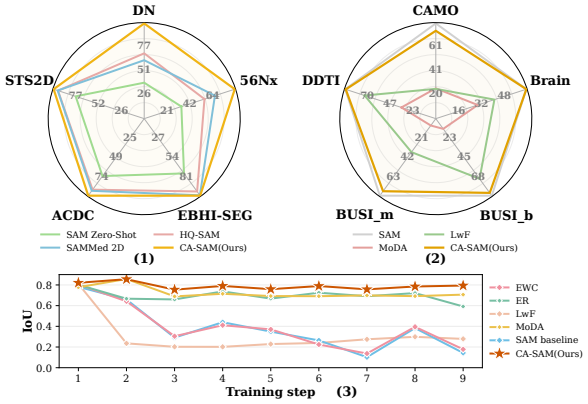


Figure 1. Comparison with prior methods across three aspects. We compare **CA-SAM (our method)** against baselines in: (1) Single-dataset segmentation performance (Radar Plot), (2) Zero-shot performance on unseen datasets (Radar Plot), and (3) Cross-dataset continual segmentation performance (Line Chart).

However, the parameter scale and computation footprint of SAM make efficient adaptation non-trivial. This bottleneck enlightens our first exploratory question: Under realistic annotation budget constraints and computation limits, do SAM-based adaptation methods offer a better cost–benefit ratio than robust CNN pipelines in practice?

Our quantitative study shows that full-parameter fine-tuning on SAM significantly outperforms CNNs but incurs a substantial increase in computational overhead. Parameter-efficient fine-tuning markedly reduces trainable parameters and computation. Nevertheless, it performs only on par with CNN pipelines [22, 38], failing to fully leverage the inherent advantages of SAM. Furthermore, in the context of Continual Learning, SAM-based adaptation methods

*Equal contribution.
†Project leader.
‡Corresponding author. Email: haixia.bi@xjtu.edu.cn

designed for downstream tasks often erode SAM's broad generalization capabilities, particularly for unseen domains.

Collectively, the evidence motivates us to rethink how SAM should be adapted for continual learning in medical image segmentation: rather than maximizing in-domain accuracy at all costs, we emphasize the importance of computational efficiency and generalization.

Concretely, we identify three limitations in current SAM-based approaches:(1) They struggle to balance computational efficiency and performance [5, 7], as they require substantial parameter updates to the backbone networks; (2) they typically depend on stored exemplars or rehearsal [28, 40], violating exemplar-free constraints common in clinical settings; (3) after fine-tuning [25, 31], they tend to erode SAM's generalization, yielding poor out-of-distribution (OOD) performance.

In response, we introduce a lightweight, plug-and-play **Alignment Layer** inserted between the frozen SAM encoder and decoder to align latent feature distributions within the target medical domain. By confining backpropagation to this module, we adapt SAM to dataset-specific statistics while preserving the capacity of the original backbone and substantially reducing computation. Building on this module, we propose **C**ontinual **A**lignment for **SAM** (**CA-SAM**), a novel continual learning framework implemented via two core components: task-specific Alignment Layers and an exemplar-free task routing mechanism. The latter, which is realized via task-calibrated Variational Autoencoders (VAEs), serves a dual purpose: it automatically adapts the Alignment Layer per task to mitigate catastrophic forgetting and falls back to the frozen SAM for zero-shot inference on unseen domains. Furthermore, CA-SAM exhibits strong robustness to task variations, a desirable virtue for real-world medical application scenarios where data patterns are inherently unpredictable. This design ensures both task-specific plasticity and SAM's broad generalization, significantly enhancing continual learning performance in medical image segmentation.

We evaluate our CA-SAM across nine medical segmentation datasets spanning multiple organs and modalities. Among SAM-based partial-fine-tuning methods [7, 16], our CA-SAM achieves state-of-the-art efficiency–performance trade-offs in computation, trainable parameters and accuracy. Under realistic streaming protocols, CA-SAM delivers strong performance with substantially reduced forgetting. Moreover, after medical-task training, the VAE Router preserves near-upper-bound zero-shot performance on previously unseen domains by correctly routing OOD inputs back to frozen SAM. Together, these results establish CA-SAM as a practical and effective recipe for continual learning in medical image segmentation with foundation models. Figure 1 presents the evaluation results of CA-SAM against other methods across three aspects: single-dataset segmentation performance, zero-shot capability and cross-dataset stability.

In summary, our contributions are as follows:

- We propose Alignment Layer: a lightweight adapter between the frozen encoder and decoder, which aligns latent representations to the medical target domain, delivering favorable computation–performance trade-offs.
- We propose CA-SAM: an exemplar-free continual learning framework with VAE-based task routing and OOD fallback to frozen SAM, mitigating forgetting while preserving zero-shot generalization.
- We conduct a comprehensive evaluation: evidence across nine heterogeneous datasets, realistic streaming settings and explicit OOD tests demonstrating state-of-the-art efficiency and accuracy.

2. Related Work

2.1. SAM for Medical Image Segmentation

The domain gap between natural and medical images is the most critical factor which limits SAM's performance in medical scenarios. Previous researchers have proposed several improvement strategies for pretrained SAM [6, 9, 23]. For example, SAM-Med2D [7], SAMed [34] and Medical SAM Adapter [31] tried to fine-tune SAM using multiple methods including LoRA and domain specific Adapter; SyncSAM [36] introduced a convolutional branch in parallel and used multi-scale feature fusion in the mask decoder; Self-Prompt SAM [32] and Semi-Supervised SAM-2 [39], respectively, employ a multi-scale prompt generator and a discriminative enhancement mechanism to enhance SAM's generalization in the medical domain. Our method, while retaining the strong generalization of pretrained SAM, efficiently performs transfer learning at the feature level using Alignment Layer with significantly fewer parameters, directly finding the best distribution for each dataset.

2.2. Segmentation Tasks in Continual Learning

Before the emergence of SAM, several studies had explored the application of continual learning in medical image segmentation. These methods can be broadly categorized into three paradigms: replay-based [3, 14, 20, 21], regularization-based [2, 18, 19], and dynamic-expansion [1, 10, 12, 33, 37] approaches. With the emergence of SAM, continual learning has entered a new phase. Using SAM strong segmentation ability and parameter-efficient fine-tuning (PEFT) techniques, recent studies pursue efficient task adaptation. Existing methods mainly include dynamic structural extensions [5, 31, 35], feature or parameter adaptation [11, 25, 26], and memory-based strategies [7, 28, 40]. However, these approaches often suffer from inefficiency, loss of task-specific information, and the lack of mechanisms to handle unseen domains, which leads to degraded

generalization of SAM. To address these limitations, we propose **CA-SAM**, a lightweight continual learning framework with learnable alignment layers and a VAE-based task routing mechanism. This design effectively preserves historical knowledge and mitigates zero-shot degradation, enabling efficient continual learning in medical image segmentation while retaining SAM's strong generalization.

3. Method

3.1. Do SAM-Based Gains Offset Resource Costs?

In recent years, SAM has been increasingly applied to medical image segmentation owing to its strong zero-shot segmentation capability. However, fine-tuning SAM, particularly full parameter fine-tuning, often entails large parameter counts and sharply increased computational costs. This raises a natural question: **are the performance gains of SAM-based methods sufficient to offset these resource costs?** To investigate this, we conduct a quantitative comparison on three medical datasets [8], contrasting CNN-baseds [22, 38] and SAM-based methods in terms of IoU/BIoU, computational cost, and the number of trainable parameters. The results are presented in Figure 2.

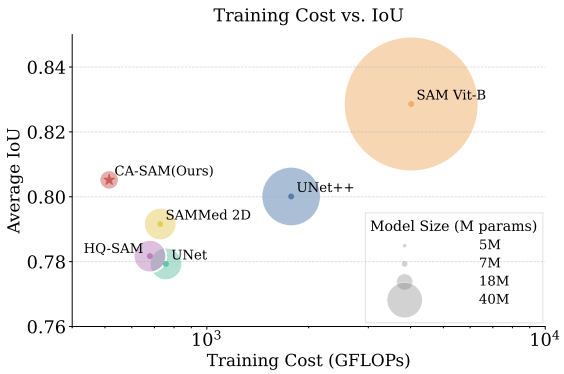


Figure 2. A comprehensive comparison of IoU scores, training cost, and trainable parameters (model size). The training computation cost is measured as the FLOPs incurred by one training pass on a standard-size image (3×1024×1024).

Our quantitative analysis shows that parameter-efficient SAM-based fine-tuning substantially reduces computational costs and the number of trainable parameters, yields clear gains in Boundary IoU(BIoU), and maintains IoU that is broadly comparable to CNN-based baselines. Meanwhile, the pronounced improvements observed with full parameter fine-tuning indicate that SAM-based approaches retain strong performance potential for medical image segmentation. Nevertheless, the parameter and computational demands of full parameter fine-tuning remain substantial. Therefore, improving fine-tuning efficiency and reducing training resource consumption are essential for the real-world deployment of SAM in this domain, which motivates our exploration of lightweight, parameter-efficient designs.

3.2. Alignment Layer

Model parameter fine-tuning methods aim to adapt feature representations across domains through distribution alignment. Since each medical dataset exhibits a distinct feature distribution, we assign a dedicated Alignment Layer to each medical dataset to align its features to an optimal distribution subspace.

Given an input image I, the structure of the pre-trained SAM can be formulated as:

$$Y = D(E(I)), \quad I \in \mathbb{R}^{B\times 3\times H\times W}, \tag{1}$$

where $E(\cdot)$ and $D(\cdot)$ denote the frozen Encoder and Decoder of SAM, respectively.

Considering the large number of parameters in the pre-trained SAM, we freeze both $E(\cdot)$ and $D(\cdot)$, and introduce a lightweight trainable module $\mathcal{A}(\cdot)$ between them to directly adapt the latent feature representation Z:

$$Z = E(I), \quad \tilde{\mathbf{Z}} = \mathcal{A}(Z). \tag{2}$$

This module $\mathcal{A}(\cdot)$ progressively approximates the target feature distribution by stacking multiple alignment layers, which are defined as several convolutional layers. The model architecture is illustrated in Section 4.1.2.

$$Y = D(\tilde{\mathbf{Z}}) = D(\mathcal{A}(E(I))), \quad Y \in \mathbb{R}^{B\times 1\times H\times W}. \tag{3}$$

This design introduces a minimal number of trainable parameters, with backpropagation restricted to the Alignment Layer, leaving the frozen Encoder and Decoder unaffected. Our method significantly reduces computational cost while preserving the architecture and capabilities of the pre-trained SAM, enabling efficient and stable domain adaptation compared to other fine-tuning approaches.

3.3. Continual Alignment for SAM

Due to privacy constraints in medical data, joint training across sources is often infeasible. To enable low-forgetting medical image segmentation under sequential training, we propose CA-SAM. Built upon a lightweight alignment layer, CA-SAM introduces an exemplar-free task routing mechanism that enables stable task identification and OOD fallback while preserving SAM's generalization ability, as illustrated in Figure 3.

Continual Learning Setting. In continual learning, training data arrive sequentially as:

$$D^{tr} = \{D_1^{tr}, \ldots, D_N^{tr}\}, \quad D^{te} = \{D_1^{te}, \ldots, D_N^{te}\}, \tag{4}$$

where D_t^{tr} and D_t^{te} denote the training and testing sets of the t-th task, and N is the total number of tasks. At

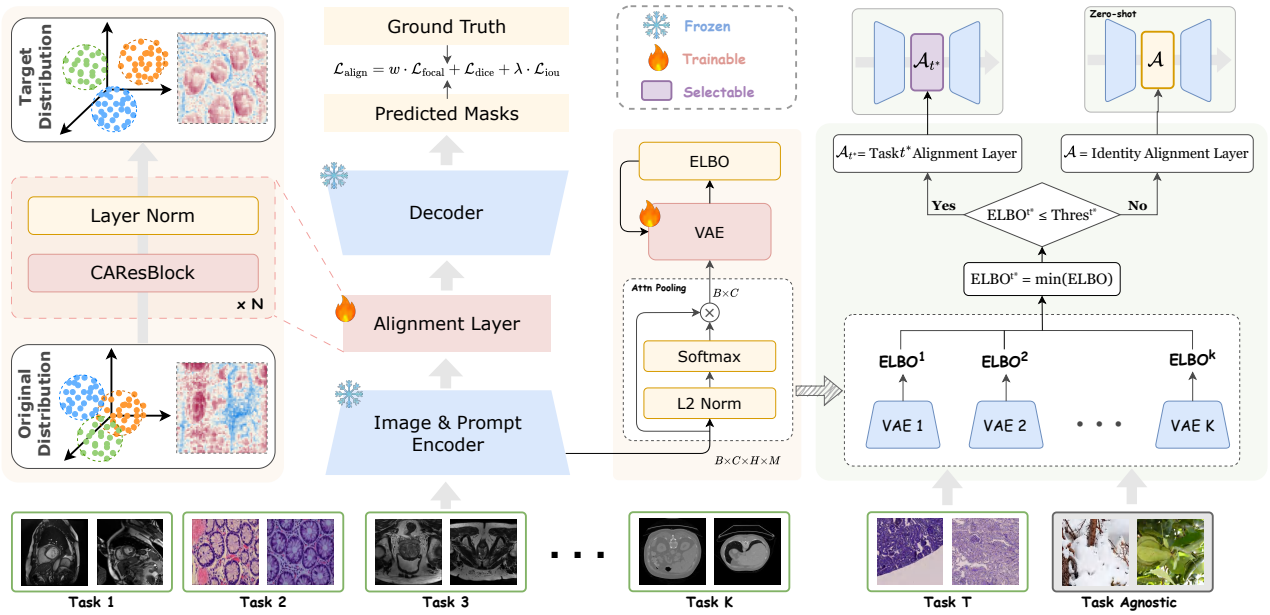


Figure 3. **Framework of Continual Alignment for SAM.** The figure shows, from left to right, the structure of the alignment layer, backbone architecture, the training procedure of the VAE, and the CA-SAM routing mechanism along with its OOD fallback mechanism.

each stage t, the model can only access the current dataset D_t^{tr} without revisiting previous samples. After training, evaluation is performed on the accumulated test sets $\{D_1^{te}, \ldots, D_t^{te}\}$ to measure both current-task performance and the degree of forgetting using IoU and BIoU metrics.

Task Routing Mechanism. For medical image segmentation tasks exhibit significant differences in feature distribution and semantic structure, we assign each task t a dedicated Alignment Layer $\mathcal{A}_t(\cdot)$ and train an independent VAE $\mathcal{V}_t = (\text{Enc}_t, \text{Dec}_t)$ to model the probabilistic feature distribution $p_t(f)$ of that task. Given the frozen encoder feature map Z, a global feature vector $f \in \mathbb{R}^D$ is obtained through a parameter-free attention pooling mechanism:

$$\alpha_{h,w} = \text{softmax}\left(\frac{\|Z_{:,h,w}\|_2}{C \cdot T}\right), f = \sum_{h,w} \alpha_{h,w} Z_{:,h,w}, \quad (5)$$

where T is a temperature coefficient, and C denotes the number of feature channels of Z. This process measures spatial saliency via the L2 norm and aggregates local features according to their importance.

Each task-specific VAE $\mathcal{V}_t$ is optimized by minimizing the evidence lower bound (ELBO) loss:

$$\mathcal{L}_{VAE}(f) = \frac{1}{D}\|f - \hat{f}\|_2^2 + \frac{\beta}{2}\sum_{i=1}^{D}(\mu_i^2 + \sigma_i^2 - 1 - \log\sigma_i^2), \quad (6)$$

where the first term is the reconstruction loss preserving feature consistency, and the second term is the Kullback–Leibler (KL) divergence regularization between the latent distribution $q_t(z|f) = \mathcal{N}(\mu, \sigma^2)$ and the standard normal prior $p(z) = \mathcal{N}(0, I)$. The coefficient β balances these two terms, where a larger β enforces stronger latent regularization, guiding the model to focus on high-level semantic distributions that better capture the inter-task differences in medical images. After optimization, each $\mathcal{V}_t$ outputs a task similarity score $s_t(f)$, which is used for task discrimination during inference.

Task Discrimination and Fallback. During inference, CA-SAM employs the trained VAEs to automatically determine the most probable task for a given input. The feature representation f is passed through all task-specific VAEs to compute ELBO-based scores:

$$s_t(f) = \mathcal{L}_{\text{VAE}}^{(t)}(f), \quad t^* = \arg\min_t s_t(f). \quad (7)$$

The task with the lowest score is regarded as the most likely match. To enhance stability, a confidence threshold τ_t is calibrated for each task via K-fold cross-validation on the training set, using the ELBO distribution of the held-out folds.

If $s_{t^*}(f) \leq \tau_{t^*}$, the model loads the corresponding alignment layer $\mathcal{A}_{t^*}$ for segmentation. Otherwise, the input is considered OOD and does not belong to any learned task. In this case, CA-SAM employs an identity alignment layer $\mathcal{A}_{id}(Z) = Z$, which reverts to the frozen SAM for zero-shot segmentation inference. This VAE Router and fallback strategy enables robust task identification and dynamic routing without explicit task labels.

4. Experiment

4.1. Experimental Settings

4.1.1. Datasets

To evaluate our method across diverse clinical settings, we use nine medical image segmentation datasets spanning organs (heart, spleen, prostate, kidney, and teeth) and modalities—magnetic resonance (MR), computed tomography (CT), and histopathology. The datasets are **ACDC** [8], **EBHI-SEG** [24], **56Nx** [27], **DN** [27], **Polyp** [15], **MSD_Prostate** [8], **MSD_Spleen** [8], **Promise12** [8], and **STS-2D** [29]. For each dataset, we follow the official train and test split.

4.1.2. Implementation Details

General Configuration. All experiments are conducted on a single NVIDIA RTX A5000 GPU (24 GB). All input images are uniformly resized to 1024×1024 to ensure compatibility with the original SAM architecture. The Adam optimizer is adopted with an initial learning rate of 1×10^{-4}, a batch size of 6 and the number of training epochs 24 for each single-task experiment. During training, random points or boxes are selected as prompts, while during testing, only box prompts are employed. The segmentation performance is evaluated on the cumulative test set of all seen tasks using two metrics: IoU and BIoU.

Alignment Layer Configuration. Specifically, our Alignment Layer consists of several stacked Channel Attention Residual Blocks (CAResBlock), each designed to enhance spatial representation and channel-wise adaptability. Each CAResBlock includes two 3×3 convolutional layers with ReLU activation, followed by an adaptive global average pooling layer. The pooled channel descriptor is further processed by a 1D convolution to model inter-channel correlations. Finally, the output feature of each CAResBlock is normalized by a LayerNorm2d layer to stabilize the feature distribution during training. During this stage, only the parameters of the Alignment Layer are trainable, while all parameters of the SAM encoder and decoder remain frozen.

Continual Learning Settings. In the continual learning experiments, we sequentially train and evaluate the model across nine medical segmentation tasks: ACDC $\rightarrow$ EBHI-SEG $\rightarrow$ 56Nx $\rightarrow$ DN $\rightarrow$ Polyp $\rightarrow$ MSD_Prostate $\rightarrow$ MSD_Spleen $\rightarrow$ Promise12 $\rightarrow$ STS-2D. Each task is assigned an independent Alignment Layer trained under the same configuration as described above. A VAE Router is introduced to achieve automatic task identification and dynamic adapter switching. Each task-specific VAE consists of a two-layer MLP encoder and decoder, with a latent dimension of 64 and a KL-divergence weight β of 16.5. The task-specific VAEs are trained for 10 epochs using a learning rate of 5×10^{-4}. For the fallback mechanism, we perform K-fold calibration ($K = 5$) to estimate the task-specific threshold τ_t, where τ_t is chosen as the p_{97} percentile of the ELBO distribution. The detailed numerical thresholds and additional implementation details for all comparison methods are provided in the Appendix.

4.2. Single-dataset Versatility Analysis

Table 3 reports results on nine medical image datasets trained and evaluated independently. For each sub-dataset, we configure a dedicated Alignment Layer, achieving the highest IoU and BIoU on the weighted average across all nine. Detailed results for every single-dataset are provided in the Appendix. Specifically, comparing to Tuning Decoder, HQ-SAM [16], and SAMMed 2D [7], our method improves the **Average IoU** by 9.75%, 7.24%, and 4.98%, respectively; and the **Average BIoU** by 12.66%, 8.11%, and 7.55%, respectively. Detailed Table in Appendix also lists the number of trainable parameters for each method. Our approach attains the best overall performance with substantially fewer trainable parameters.

4.3. Cross-dataset Stability Details

Quantitative Evaluation. Table 1 presents the quantitative results under our experimental setup. The key observations are summarized as follows: (1) Naive adapter finetuning suffers from severe forgetting. Its performance is far below the joint-training upper bound. This indicates that naive adapter updates cannot resist cross-task distributional shifts. (2) Classical continual learning methods such as LwF, EWC, ER, and DER moderately mitigate forgetting, demonstrating the necessity of introducing continual learning mechanisms. However, their effectiveness remains limited under pixel-level medical image segmentation with strong distributional shifts. (3) Despite their parameter efficiency, the prompt-based method shows limited gains. (4) Parameter-fusion methods show limitations when applied to multi-task learning involving cross-modality and cross-organ data. (5) MoDA greatly outperforms other CL methods. However, CA-SAM further achieves the highest performance under the same frozen SAM setting, reducing forgetting to nearly zero. Compared with the second-best approach, CA-SAM improves *Last-IoU* and *Last-BIoU* by 5.48% and 4.43%, respectively. Notably, methods that use routing exhibit strong robustness (see Appendix) and are inherently invariant to order, whereas methods such as EWC and LwF tend to suffer from catastrophic forgetting when consecutive tasks differ substantially in modality.

Visual Evaluation. Figure 4 presents the visual results and IoU scores of different methods after completing continual training on all nine tasks. Continual learning methods such as ER, EWC, and LwF still suffer from forgetting, and their performance even falls below SAM on several datasets.In contrast, our method delivers accurate and stable segmentation across all tasks, showing significant improvements in accuracy and stability over SAM. Even on challenging

Method	EF	IoU on Med			BIoU on Med		
		Last-IoU ↑	Avg-IoU ↑	FF-IoU ↓	Last-BIoU ↑	Avg-BIoU ↑	FF-BIoU ↓
Joint training	✓	78.43	—	—	61.26	—	—
SAM [17]	✓	53.81	—	—	36.52	—	—
SAM [17] + AL(naive)	✓	14.38	37.93	65.79%	13.50	29.91	49.69%
LwF [19]	✓	27.84	30.77	3.03%	22.82	28.45	3.05%
EWC [18]	✓	17.78	38.51	57.32%	12.58	30.42	42.16%
EMR [13]	✓	43.09	53.96	10.95%	26.05	33.77	8.53%
ER [21]	✗	59.19	69.59	21.68%	39.93	51.37	24.25%
DER [4]	✗	59.99	63.14	6.55%	40.69	42.20	6.69%
L2P [30]	✓	57.34	53.84	1.55%	37.86	35.70	1.79%
MoDA [35] + HQ-SAM [16]	✗	69.30	70.86	2.00%	55.52	55.88	1.64%
MoDA [35]	✗	70.64	72.44	0.66%	52.12	54.45	0.65%
CA-SAM (Ours)	✓	**76.12**	**76.90**	**1.43%**	**59.95**	**59.45**	**0.24%**

Table 1. **Continual Learning results on medical datasets.** "EF" indicates Exemplar-free continual learning methods that do not access previous task samples during current task training. All compared methods adopt an Alignment Layer for SAM adaptation, except *MoDA + HQ-SAM*, which uses the HQ-SAM decoder. Details of metrics are in Appendix. The best and second best performances are highlighted.

Method	DDTI		BUSI_benign		BUSI_malignant		Brain_Tumor		CAMO	
	IoU	BIoU	IoU	BIoU	IoU	BIoU	IoU	BIoU	IoU	BIoU
SAM [17]	78.08	45.98	75.73	53.30	69.90	33.70	52.85	60.89	67.60	43.05
MoDA [35]	29.66 (48.42)	20.18 (25.80)	10.06 (65.67)	11.45 (41.85)	6.24 (63.66)	6.50 (27.20)	24.47 (28.38)	30.59 (30.30)	21.51 (46.09)	13.85 (29.20)
LwF [19]	61.34 (16.74)	21.22 (24.76)	59.16 (16.57)	37.64 (15.66)	30.38 (39.52)	13.07 (20.63)	34.20 (18.65)	27.46 (33.43)	21.14 (46.46)	12.79 (30.26)
CA-SAM	77.62 (**0.46**)	45.73 (**0.25**)	72.99 (**2.74**)	51.67 (**1.63**)	65.86 (**4.04**)	31.66 (**2.04**)	52.85 (**0.00**)	60.89 (**0.00**)	62.37 (**5.23**)	39.60 (**3.45**)

Table 2. Zero-shot results of Continual Learning methods. **Green** and **red** numbers denote the decrease and increase relative to SAM.

Methods	Parameters	GFLOPs	Avg-IoU	Avg-BIoU
SAM [17]	–	–	55.08	37.67
Tuning Decoder [17]	4.06M	669.8	70.40	53.86
HQ-SAM [16]	5.14M	678.9	72.91	58.41
SAMMed 2D [7]	13.31M	728.2	75.17	58.97
CA-SAM (Ours)	**3.54M**	**514.3**	**80.15**	**66.52**

Table 3. Single-dataset Versatility Results.

tasks (e.g., T2 and T4), where other methods fail and SAM produces incorrect masks, our approach still produces high-quality and coherent segmentation results.

4.4. Evaluation on Out-of-Distribution Data

To assess whether continual learning compromises the generalization ability of SAM, we evaluate each model's zero-shot segmentation performance on unseen domains after completing training on all medical tasks. All comparison methods are reproduced using their publicly available implementations, and the results are reported in Table 2. Due to the lack of effective mechanisms for preserving generalization, models trained with LwF and MoDA suffer from a severe loss of SAM's zero-shot capability. In particular, MoDA's hard routing often misassigns OOD samples to incorrect adapters, causing feature mismatch and eroding the base model's general knowledge. In contrast, our proposed *CA-SAM* introduces a VAE-based task discrimination and OOD fallback mechanism that maximally preserves SAM's inherent generalization. Experimental results on five unseen datasets demonstrate that the VAE Router achieves a weighted average discrimination accuracy of 99.42% (DDTI: 99.37%, BUSI_benign: 96.34%, BUSI_malignant: 92.38%, Brain_Tumor: 100.00%, CAMO: 92.00%). Most OOD samples are correctly routed back to SAM for zero-shot inference, enabling our method to achieve zero-shot performance on unseen domains that is nearly equivalent to the SAM upper bound.

4.5. Explainability

To verify whether the Alignment Layer can align the feature distributions of different datasets to their respective optimal distribution spaces, we conducted experiments using

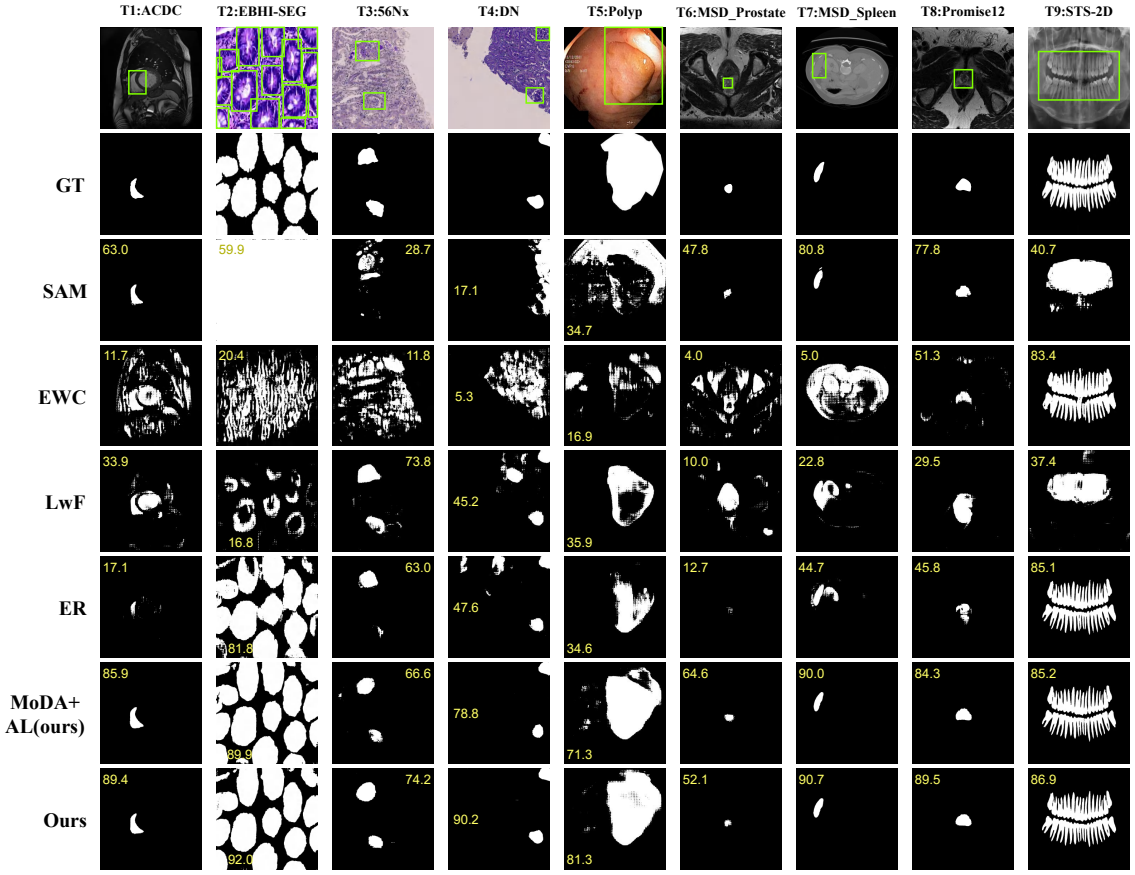


Figure 4. Qualitative comparison of segmentation results and corresponding IoU scores after continual training on all nine tasks. The figure illustrates the performance differences among competing continual learning methods. AL denotes the proposed Alignment Layer module.

56Nx	TV Distance ↓	JS Divergence ↓
Before vs Adapter	0.2830	0.2398
After vs Adapter	0.2742	0.2351

DN	TV Distance ↓	JS Divergence ↓
Before vs Adapter	0.2840	0.2404
After vs Adapter	0.2832	0.2365

Table 4. Feature Distribution Distance Comparison in 56Nx, DN

the 56Nx and DN datasets. We compared the feature maps output by the SAM Encoder after Adapter fine-tuning [31] as a baseline, and calculated the distance between the features before and after the Alignment Layer and the baseline features. We selected the TV (Total Variation) and JS (Jensen-Shannon) divergence metrics for comparison. See the Appendix for details on the metric formulations.

The results are shown in the Table 4 and the Figure 5. The experiments show that the feature maps after the Alignment Layer are closer to the outputs of the SAM Encoder tuning than those before the Alignment Layer, confirming that the Alignment Layer effectively aligns the features within the target medical domain.

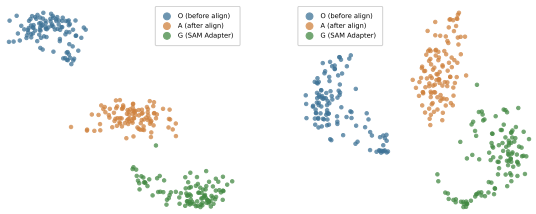


Figure 5. Two TSNE visualization comparison images: the left image is 56Nx dataset, and the right image is DN dataset.

4.6. Ablation Study

4.6.1. Single-dataset Versatility Ablation

We experimented with different numbers of Blocks in Alignment Layer, and collected the average IoU/BIoU

Method	Hyper-Param	IoU on Med			BIoU on Med		
		Last-IoU	Avg-IoU	FF-IoU	Last-BIoU	Avg-BIoU	FF-BIoU
Global Average Pooling	$\beta = 16.5$	75.93	76.87	1.65	59.63	59.38	0.42
Mean	$\beta = 16.5$	75.86	76.91	1.61	59.72	59.46	0.38
Flatten	$\beta = 16.5$	24.06	33.64	7.55	20.95	28.94	6.34
CLS Token	$\beta = 16.5$	49.71	48.19	0.01	31.23	28.42	0.02
Attn Pooling (With Params)	$\beta = 16.5$, temp = 1	71.03	72.59	1.14	56.39	56.52	0.71
Attn Pooling (No Params)	$\beta = 16.5$, temp = 1	**76.41**	**77.07**	1.28	**59.91**	**59.51**	0.27

Table 5. Comparison of different pooling methods and their performance metrics.

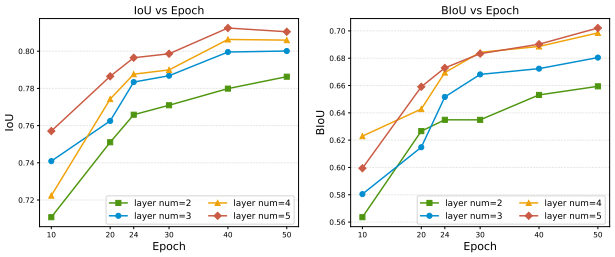


Figure 6. Ablation Study with Number of Alignment Layers.

across different epochs. The experimental results are shown in the Figure 6. We found that, increasing the number of Alignment Layers can enhance its fitting ability, thus correspondingly improving the performance. The results demonstrate the effectiveness of proposed CA-SAM.

4.6.2. Cross-dataset Stability Ablation

Feature Pooling Ablation. We evaluate multiple feature-pooling designs and report IoU/BIoU after sequential training and testing on nine medical datasets, as summarized in the Table 5. The candidates include GAP (global average pooling over spatial dimensions of the encoder feature map), Mean pooling, Flatten (reshape a 2D feature map into a vector followed by a linear projection to the target dimension), CLS token (take the Transformer's [CLS] output directly), parameter-free attention pooling (compute the ℓ_2 norm per spatial location on the feature map, normalize with Softmax to obtain attention over (H, W), and use it to weighted-sum the original (B, C, H, W) features into (B, C)), and learnable attention pooling (as above, but with trainable parameters to generate attention weights). Across all settings, attention pooling yields the strongest segmentation performance, consistently outperforming the alternatives by better focusing on salient regions and providing the VAE with more discriminative global features. Notably, the parameter-free attention pooling achieves the best task separability without introducing any additional parameters, delivering the overall best results among the tested strategies.

Router Threshold (τ) Ablation. The router threshold τ is the key hyperparameter for rejecting unknown tasks. To improve decision stability, we calibrate a separate threshold τ_t for each task t. For this calibration, we perform K-fold cross-validation on the training set using the task's VAE V_t. We evaluate several rules for setting τ_t based on the ELBO score distribution on the training set, namely $\mu + 2\sigma$ and the p_{95}/p_{97}/p_{99} percentiles.The results in Table 6 reveal a trade-off between *known-task routing accuracy* and *zero-shot (OOD) performance*. A lenient threshold such as p_{99} biases the router toward "known," which improves seen-task continual segmentation (IoU on Med 76.04%) but degrades OOD discrimination (Acc 98.76%). In contrast, p_{97} applies a tighter cutoff, yielding stronger OOD accuracy (Acc 99.42%) *without material loss* on seen-task continual segmentation. We therefore adopt p_{97} as our calibration rule, striking a better balance between OOD detection and segmentation on known tasks.

Parameter	IoU (Med)	BIoU (Med)	IoU	BIoU	Acc
$\mu + 2\sigma$	75.84	59.66	55.87	58.27	99.33
p_{95}	75.83	59.56	**56.08**	**58.39**	**99.66**
p_{97}	76.02	59.64	55.92	58.29	99.42
p_{99}	**76.04**	**59.71**	55.48	58.08	98.76

Table 6. **Comparison of CA-SAM configurations under different τ thresholds.** "Acc" denotes OOD discrimination accuracy.

5. Conclusion

In this paper, we demonstrate that SAM provides strong segmentation priors and considerable potential for continual learning. However, fine-tuning all parameters incurs considerable computational and parameter overheads. To balance performance and efficiency, we introduce a lightweight and plug-and-play Alignment Layer that aligns features between the encoder and decoder, significantly reducing trainable parameters while maintaining high segmentation accuracy. Building on the Alignment Layer, we propose Continual Alignment for SAM (CA-SAM), an exemplar-free continual learning framework that automatically identifies tasks and routes inputs to the appropriate Alignment Layer or the

frozen SAM for OOD fallback. By providing robustness to varying task orders and strong performance across diverse medical datasets, CA-SAM highlights a promising direction for enabling scalable and reliable continual learning of pre-trained foundation models in medical image segmentation.

References

[1] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3366–3375, 2017. 2

[2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018. 2

[3] Sutanu Bera, Vinay Ummadi, Debashis Sen, Subhamoy Mandal, and Prabir Kumar Biswas. Memory replay for continual medical image segmentation through atypical sample selection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 513–522. Springer, 2023. 2

[4] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline, 2020. 6

[5] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3367–3375, 2023. 2

[6] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. Sam on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*, 2023. 2

[7] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiangand Hui Sun, Junjun He, Shaoting Zhang, Min Zhu, and Yu Qiao. Sam-med2d, 2023. 1, 2, 5, 6, 3

[8] Junlong Cheng, Bin Fu, Jin Ye, Guoan Wang, Tianbin Li, Haoyu Wang, Ruoyu Li, He Yao, Junren Cheng, JingWen Li, et al. Interactive medical image segmentation: A benchmark dataset and baseline. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20841–20851, 2025. 3, 5

[9] Zhiheng Cheng, Qingyue Wei, Hongru Zhu, Yan Wang, Liangqiong Qu, Wei Shao, and Yuyin Zhou. Unleashing the potential of sam for medical adaptation via hierarchical decoding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3511–3522, 2024. 2

[10] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9285–9295, 2022. 2

[11] Takuma Fukuda, Hiroshi Kera, and Kazuhiko Kawamoto. Adapter merging with centroid prototype mapping for scalable class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4884–4893, 2025. 2

[12] Zhiyuan Hu, Yunsheng Li, Jiancheng Lyu, Dashan Gao, and Nuno Vasconcelos. Dense network expansion for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11858–11867, 2023. 2

[13] Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. Emr-merging: Tuning-free high-performance model merging. *Advances in Neural Information Processing Systems*, 37:122741–122769, 2024. 6

[14] David Isele and Akansel Cosgun. Selective experience replay for lifelong learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[15] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019. 5

[16] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 2, 5, 6, 3

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1, 6, 3, 5

[18] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2, 6, 5

[19] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 2, 6, 5

[20] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 2

[21] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019. 2, 6, 5

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 3

[23] Tal Shaharabany, Aviad Dahan, Raja Giryes, and Lior Wolf. Autosam: Adapting sam to medical images by overloading the prompt encoder. *arXiv preprint arXiv:2306.06370*, 2023. 2

[24] Liyu Shi, Xiaoyan Li, Weiming Hu, Haoyuan Chen, Jing Chen, Zizhen Fan, Minghe Gao, Yujie Jing, Guotao Lu,

Deguo Ma, et al. Ebhi-seg: A novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. *Frontiers in Medicine*, 10: 1114673, 2023. 5

[25] Weili Shi, Penglong Zhang, Yuqin Li, and Zhengang Jiang. Segment anything model for few-shot medical image segmentation with domain tuning. *Complex & Intelligent Systems*, 11(1):37, 2025. 2

[26] Yuan-Chen Shu, Zhiwei Lin, and Yongtao Wang. Regcl: Continual adaptation of segment anything model via model merging. *arXiv preprint arXiv:2507.12297*, 2025. 2

[27] Yucheng Tang, Yufan He, Vishwesh Nath, Pengfeig Guo, Ruining Deng, Tianyuan Yao, Quan Liu, Can Cui, Mengmeng Yin, Ziyue Xu, et al. Holohisto: End-to-end gigapixel wsi segmentation with 4k resolution sequential tokenization. *arXiv preprint arXiv:2407.03307*, 2024. 5

[28] Guankun Wang, Long Bai, Yanan Wu, Tong Chen, and Hongliang Ren. Rethinking exemplars for continual semantic segmentation in endoscopy scenes: Entropy-based mini-batch pseudo-replay. *Computers in Biology and Medicine*, 165:107412, 2023. 2

[29] Yaqi Wang, Fan Ye, Yifei Chen, Chengkai Wang, Chengyu Wu, Feng Xu, Zhean Ma, Yi Liu, Yifan Zhang, Mingguo Cao, et al. A multi-modal dental dataset for semi-supervised deep learning image segmentation. *Scientific Data*, 12(1): 117, 2025. 5

[30] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022. 6, 5

[31] Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547, 2025. 2, 7

[32] Bin Xie, Hao Tang, Dawen Cai, Yan Yan, and Gady Agam. Self-prompt sam: Medical image segmentation via automatic prompt sam adaptation. *arXiv preprint arXiv:2502.00630*, 2025. 2

[33] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3014–3023, 2021. 2

[34] Zhiling Yan, Sifan Song, Dingjie Song, Yiwei Li, Rong Zhou, Weixiang Sun, Zhennong Chen, Sekeun Kim, Hui Ren, Tianming Liu, et al. Samed-2: Selective memory enhanced medical segment anything model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–550. Springer, 2025. 2

[35] Jinglong Yang, Yichen Wu, Jun Cen, Wenjian Huang, Hong Wang, and Jianguo Zhang. Continual learning for segment anything model adaptation. *arXiv preprint arXiv:2412.06418*, 2024. 1, 2, 6, 5

[36] Sihan Yang, Jiadong Feng, Xuande Mi, Haixia Bi, Hai Zhang, and Jian Sun. Improved baselines with synchronized encoding for universal medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 260–270. Springer, 2025. 2

[37] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017. 2

[38] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pages 3–11. Springer, 2018. 1, 3

[39] Hongjie Zhu, Xiwei Liu, Rundong Xue, Zeyu Zhang, Yong Xu, Daji Ergu, Ying Cai, and Yang Zhao. Sss: Semi-supervised sam-2 with efficient prompting for medical imaging segmentation. *arXiv preprint arXiv:2506.08949*, 2025. 2

[40] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Continual semantic segmentation with automatic memory sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3082–3092, 2023. 1, 2

Continual Alignment for SAM: Rethinking Foundation Models for Medical Image Segmentation in Continual Learning

Supplementary Material

A. Dataset Details

We selected nine medical datasets, each representing different anatomical regions and imaging modalities. Detailed information about these datasets is provided below.

ACDC: The ACDC dataset consists of 1632 training images and 177 test images. The dataset modality is MR (Magnetic Resonance) imaging. The segmentation targets include three parts of the heart: the left ventricle, myocardium, and right ventricle.

EBHI-SEG: The EBHI-SEG dataset contains 1701 training images and 487 test images. The dataset modality is pathological imaging, with the segmentation target being colon cancer (affected areas).

56Nx: The 56Nx dataset includes 558 training images and 463 test images. The dataset modality is pathological imaging, with the segmentation target being the glomerulus.

DN: The DN dataset contains 724 training images and 391 test images. The dataset modality is pathological imaging, and the segmentation target is the glomerulus.

Ployp: The Ployp dataset consists of 804 training images and 196 test images. The dataset modality is RGB imaging, and the segmentation target is the spleen.

MSD_prostate: The MSD_prostate dataset includes 419 training images and 53 test images. The dataset modality is MR T2 (Magnetic Resonance Imaging), with the segmentation target being two regions of the prostate: the peripheral zone and transition zone.

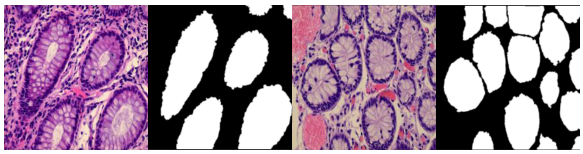
MSD_Spleen: The MSD_Spleen dataset contains 876 training images and 146 test images. The dataset modality is CT imaging, and the segmentation target is the spleen.

promise12: The promise12 dataset includes 712 training images and 66 test images. The dataset modality is MR (Magnetic Resonance) imaging, and the segmentation target is the prostate.

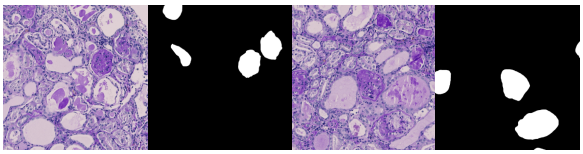
STS-2D: The STS-2D dataset consists of 1700 training images and 70 test images. The dataset modality is X-ray imaging, and the segmentation target is the teeth.

Figure 7 present representative examples of the images and corresponding masks for each dataset.

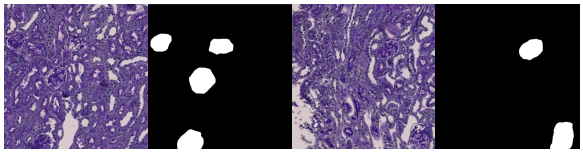
(1) ACDC



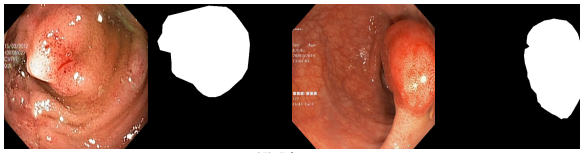
(2) EBHI-SEG



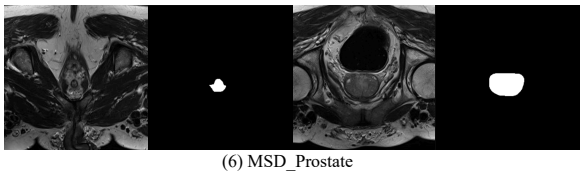
(3) 56Nx



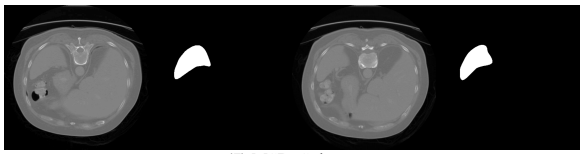
(4) DN



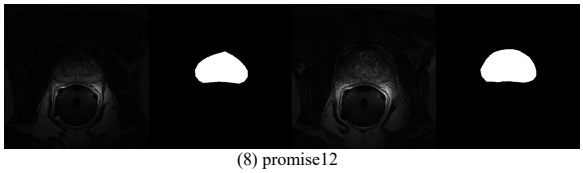
(5) Ployp



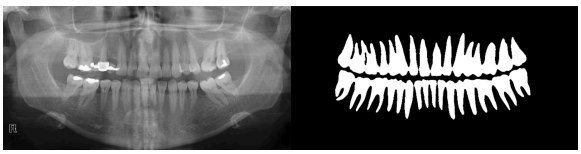
(6) MSD_Prostate



(7) MSD_Spleen



(8) promise12



(9) STS-2D

Figure 7. **Datasets of Task 1-9 :** (1)-(9) show the examples of the images and corresponding masks for nine medical dataset.

B. Experiment Evaluation Metrics

In the Cross-dataset Stability experiment, we selected three metrics: **Last-IoU**, **Avg-IoU** and **FF-IoU** to evaluate the continual segmentation performance of different methods across nine datasets. To evaluate the final overall accuracy, we employ the metrics Last-IoU and Last-BIoU to measure the segmentation performance across all tasks after the completion of all sequential tasks. We define the average performance after training t tasks, denoted as IoU_t,

$$\text{IoU}_t = \frac{1}{N_t}\sum_{k=1}^{t} n_k \,\text{IoU}_{k,t}, \quad \text{BIoU}_t = \frac{1}{N_t}\sum_{k=1}^{t} n_k \,\text{BIoU}_{k,t}, \tag{8}$$

where $\text{IoU}_{k,t}$ and $\text{BIoU}_{k,t}$ represent the weighted IoU/BIoU evaluated on the test set of the k-th task after training on t tasks, and n_k denotes the number of images in the test set of the k-th task and $N_t = \sum_{k=1}^{t} n_k$. Then, the Last-IoU and Last-BIoU are defined as

$$\text{Last-IoU} = \text{IoU}_N, \quad \text{Last-BIoU} = \text{BIoU}_N. \tag{9}$$

To illustrate the average segmentation performance throughout the training process during sequential learning, we also use the Avg-IoU and Avg-BIoU metrics, as described below:

$$\text{Avg-IoU} = \frac{1}{N}\sum_{t=1}^{N} \text{IoU}_t, \quad \text{Avg-BIoU} = \frac{1}{N}\sum_{t=1}^{N} \text{BIoU}_t. \tag{10}$$

To measure forgetting performance, we use $f_{k,t}$ as the forgetting on task t after training on all t tasks,

$$f_{k,t} = \max_{j\in\{1,\ldots,t-1\}} \left(\text{IoU}_{k,j} - \text{IoU}_{k,t}\right). \tag{11}$$

Then, the average forgetting measure, defined as FF-IoU, can be computed after training on all *N* tasks.

$$\text{FF-IoU} = \frac{1}{N-1}\sum_{k=1}^{N-1} f_{k,N}. \tag{12}$$

The metric FF-BIoU is calculated in a similar way.

C. Single-dataset Versatility Analysis

Table 7 in the Appendix reports the complete single-dataset experiment results for training and testing the proposed *Alignment Layer* on each of the nine medical datasets. Specifically, our method attains the best performance on five out of nine datasets (ACDC, EBHI-SEG, 56Nx, DN and STS-2D) as well as on the weighted average across datasets. Notably, on *56Nx* and *DN* we surpass the second-best approach by **10.3%** and **24.9%** in IoU, and by **9.2%** and **28.0%** in BIoU, respectively. Taken together, these results provide strong evidence for the effectiveness of the proposed Alignment Layer.

D. Cross-dataset Stability Details

D.1. Pipeline of Algorithms

To assess the capability of contrastive methods for medical image segmentation in continual learning, we adapt and modify each method for the alignment-based SAM framework as follows:

1) For **LwF**, the alignment layers trained on the previous task act as the teacher network, and a knowledge distillation loss constrains the current alignment layers to retain consistent outputs with the teacher while learning new knowledge, thereby mitigating catastrophic forgetting.

2) For **EWC**, we estimate the importance of each parameter using the Fisher Information Matrix and penalize changes to crucial parameters during new task training to preserve previously learned knowledge.

3) For **ER**, a memory bank stores a small set of samples from past tasks, which are replayed and jointly trained with new task data to maintain prior performance.

4) For **DER**, this method further introduces knowledge distillation on the basis of ER. It constrains the consistency between the current model and historical model predictions by storing and replaying past samples in memory and matching the network's Logits sampled throughout the optimization trajectory.

5)For **L2P**, we maintain a cumulative prompt pool that uses SAM image embeddings as query keys to retrieve the most relevant prompts, while a slot-based allocation mechanism ensures task-wise isolation and efficient prompt utilization.

6)For **MoDA**, we introduce a task classifier by augmenting the encoder with a [CLS] token that captures global task information; during inference, the classifier automatically routes the input image to its corresponding historical alignment layer according to the [CLS] token.

7)For **EMR**, we employ a parameter space merging strategy on the alignment layer. This method defines the task vectors τ_i as the weight increments from each historical task. These vectors are then aggregated into a single unified task vector τ_{uni} via an Electing procedure, which selects the parameter with the largest magnitude that is consistent in sign across all tasks. During inference, τ_{uni} is subject to task-specific modulation: an alignment mask (M_t) is applied to filter τ_{uni} by zeroing out parameters that conflict in direction with the specific task vector, and a rescaling factor λ_t is used to calibrate the magnitude of the modulated vector to match the original task vector's scale, thereby allowing the single layer to efficiently store and recall knowledge for each specific task.

8)We further include **Joint Training** as an upper bound, where all task datasets are trained simultaneously.

9)**Naive approach** as a lower bound. This method is a simple Sequential Fine-Tuning baseline under the Frozen SAM framework, where only the alignment layers are se-

Methods(Iou)	Parameters	ACDC	EBHI_SEG	56Nx	DN	Polyp	MSD_Prostate	MSD_Spleen	promise12	STS2D	Average Iou
Zero-Shot [17]	0 M	61.47	63.98	29.06	32.12	66.61	60.55	81.39	83.23	65.10	55.08
Tuning Decoder	4.06M	68.45	89.53	43.05	50.96	81.49	67.42	**91.98**	**90.58**	82.38	70.40
HQ-SAM [16]	5.14M	77.05	85.07	47.11	58.34	78.46	**72.33**	90.27	87.65	81.88	72.91
SAMMed 2D [7]	13.31M	78.27	**89.58**	55.29	52.20	**86.87**	71.46	91.72	86.69	82.16	75.17
CA-SAM(Ours)	**3.54M**	**80.75**	89.14	**65.92**	**83.23**	65.02	62.71	86.18	84.52	**86.03**	**80.15**

Methods(BIou)	Parameters	ACDC	EBHI_SEG	56Nx	DN	Polyp	MSD_Prostate	MSD_Spleen	promise12	STS2D	Average BIou
Zero-Shot [17]	0 M	57.83	19.85	22.17	22.63	42.48	53.11	75.44	64.42	36.45	37.67
Tuning Decoder	4.06M	65.44	54.09	22.28	29.41	55.52	64.55	88.00	**77.56**	80.85	53.86
HQ-SAM [16]	5.14M	73.18	50.18	33.46	43.04	56.06	**65.24**	86.52	71.97	79.63	58.41
SAMMed 2D [7]	13.31M	73.09	54.44	35.99	31.71	**65.65**	64.88	**89.22**	73.93	80.84	58.97
CA-SAM(Ours)	**3.54M**	**75.90**	**62.93**	**45.18**	**71.01**	42.16	53.62	87.19	69.52	**84.75**	**66.52**

Table 7. Detailed Single Med-Dataset Versatility Results

quentially updated without any CL mechanism to quantify cross-task forgetting.

Algorithm 1. Continual Alignment for SAM with LwF

Input: Pre-trained SAM θ, Tasks $\mathcal{D} = \{D_1^{tr}, \ldots, D_N^{tr}\}$
Output: $\mathcal{A}_N$
Initialize $\mathcal{A}_1$ and train on D_1^{tr}.
for $t = 2, \ldots, N$ **do**
  Set teacher $\mathcal{A}_{t-1}$ and initialize $\mathcal{A}_t \leftarrow \mathcal{A}_{t-1}$.
  for $(x, y) \in D_t^{tr}$ **do**
    $\hat{m}_t = f_\theta(x; \mathcal{A}_t)$, $\hat{m}_{t-1} = f_\theta(x; \mathcal{A}_{t-1})$.
    Update $\mathcal{A}_t$ with $\mathcal{L} = \mathcal{L}_{align}(\hat{m}_t, y) + \lambda \mathcal{L}_{mask}(\hat{m}_t, \hat{m}_{t-1})$.
  end for
end for
return $\mathcal{A}_N$

Algorithm 2. Continual Alignment for SAM with EWC

Input: Pre-trained SAM θ, Tasks $\mathcal{D} = \{D_1^{tr}, \ldots, D_N^{tr}\}$
Output: $\mathcal{A}_N$, $\mathcal{F}_N$
Train $\mathcal{A}_1$ on D_1^{tr} and compute Fisher $\mathcal{F}_1$.
for $t = 2, \ldots, N$ **do**
  Initialize $\mathcal{A}_t \leftarrow \mathcal{A}_{t-1}$.
  for $(x, y) \in D_t^{tr}$ **do**
    Compute $\mathcal{L}_{align}$ on D_t^{tr}
    Compute $\mathcal{L}_{ewc} = \sum_i \mathcal{F}_{t-1,i}(\mathcal{A}_{t,i} - \mathcal{A}_{t-1,i})^2$.
    Update $\mathcal{A}_t$ using $\mathcal{L} = \mathcal{L}_{align} + \lambda \mathcal{L}_{ewc}$.
  end for
  Estimate new $\mathcal{F}_t$.
end for
return $\mathcal{A}_N$, $\mathcal{F}_N$

Algorithm 3. Continual Alignment for SAM with ER

Input: Pre-trained SAM θ, Tasks $\mathcal{D} = \{D_1^{tr}, \ldots, D_N^{tr}\}$
Output: $\mathcal{A}_N$, $\mathcal{M}$
Initialize Memory Bank $\mathcal{M} \leftarrow \emptyset$.
for $t = 1, \ldots, N$ **do**
  initialize $\mathcal{A}_t \leftarrow \mathcal{A}_{t-1}$
  for $(x, y) \in D_t^{tr} \cup \mathcal{M}$ **do**
    Compute $\mathcal{L}_{align}$ on D_t^{tr}
    Update $\mathcal{A}_t$ with loss $\mathcal{L}_{align}$.
  end for
  Select exemplars from D_t^{tr} to $\mathcal{M}$.
end for
return $\mathcal{A}_N$, $\mathcal{M}$

Algorithm 4. Continual Alignment for SAM with DER

Input: Pre-trained SAM θ, Tasks $\mathcal{D} = \{D_1^{tr}, \ldots, D_N^{tr}\}$
Output: $\mathcal{A}_N$, $\mathcal{M}$
Initialize Memory Bank $\mathcal{M} \leftarrow \emptyset$.
for $t = 1, \ldots, N$ **do**
  initialize $\mathcal{A}_t \leftarrow \mathcal{A}_{t-1}$
  for $(x, y) \in D_t^{tr}$ **do**
    $(x', z') \leftarrow$ Sample($\mathcal{M}$)
    $Z_t = E_\theta(x)$; $\tilde{Z}_t = \mathcal{A}_t(Z_t)$; $h_t = D_\theta(\tilde{Z}_t)$
    $\mathcal{L}_{new} = \mathcal{L}_{align}(y, h_t)$
    $Z' = E_\theta(x')$; $\tilde{Z}' = \mathcal{A}_t(Z')$; $h' = D_\theta(\tilde{Z}')$
    $\mathcal{L}_{distill} = ||z' - h'||_2^2$
    $\mathcal{L}_{total} = \mathcal{L}_{new} + \alpha \mathcal{L}_{distill}$
    Update $\mathcal{A}_t$ with loss $\mathcal{L}_{total}$.
    $\mathcal{M} \leftarrow$ ReservoirUpdate($\mathcal{M}$, (x, h_t))
  end for
end for
return $\mathcal{A}_N$, $\mathcal{M}$

D.2. Evaluation on Task-Order Robustness

According to Table 8, different continual learning methods exhibit clear variations in performance when the task order changes. Traditional continual learning approaches such as LwF, EWC, and naive adapter fine-tuning show large fluctuations in Last-IoU, Avg-IoU, and forgetting across different orders, indicating strong sensitivity to cross-task distribution shifts and a lack of robustness to task ordering. ER is relatively more stable, but its performance is still affected by task permutations.

In contrast, both MoDA and CA-SAM maintain consistently high performance under all three task orders. Since these routing-based methods activate separate adapter parameters for each task, they naturally avoid interference be-

Algorithm 5. Continual Alignment for SAM with L2P

Input: Pre-trained SAM θ, Pre-trained Alignment layer $\mathcal{A}$
Tasks $\mathcal{D} = \{D_1^{tr}, \ldots, D_N^{tr}\}$
Output: Prompt Pool $\mathcal{P}_N$
Initialize Prompt Pool $\mathcal{P}$.
If using task slots: assign each task its prompt range.
for $t = 1, \ldots, N$ **do**
  for $(x, y) \in D_t^{tr}$ **do**
    $Z = E_\theta(x)$; $\tilde{Z} = \mathcal{A}_t(Z)$.
    retrieve top-k prompts p_k from $\mathcal{P}$.
    $\hat{m} = D_\theta(\tilde{Z}; p_k)$.
    Compute $\mathcal{L}_{align}(\hat{m}, y)$ and $\mathcal{L}_{key-match}$.
    Update $\mathcal{P}$.
  end for
  Save $\mathcal{P}_t$ for next task.
end for
return $\mathcal{P}_N$

Algorithm 6. Continual Alignment for SAM with MoDA

Input: Pre-trained SAM θ, Tasks $\mathcal{D} = \{D_1^{tr}, \ldots, D_N^{tr}\}$
Output: Alignment Layer Pool $\mathcal{P} = \{K_t : \Phi_t\}_{t=1}^N$, Task Classifier $\mathcal{T}$ (with global tokens $T \in \mathbb{R}^{L \times C}$)
Initialize $\mathcal{P} \leftarrow \emptyset$, global tokens T, router $\mathcal{T}$.
for $t = 1, \ldots, N$ **do**
  Initialize / load current alignment layer Φ_t.
  for $(x, y) \in D_t^{tr}$ **do**
    $Z = E_\theta(x)$; $\tilde{Z} = \Phi_t(Z)$; $\hat{m} = f_\theta(\tilde{Z})$
    Update Φ_t by $\mathcal{L}_{align}(\hat{m}, y)$.
  end for
  Save Φ_t into pool: $\mathcal{P} \leftarrow \mathcal{P} \cup \{K_t : \Phi_t\}$.
  Select exemplars for Memory Bank $\mathcal{M}$.
  for $x \in \mathcal{M}$ **do**
    global feature $q = f'_\theta(T)[0]$.
    Update router $\mathcal{T}$ with CE loss $\mathcal{L}_{ce}$.
  end for
end for
return $\mathcal{P}$, $\mathcal{T}$

tween tasks and are therefore largely invariant to task order. Notably, CA-SAM achieves the highest Last-IoU and Avg-IoU as well as the lowest forgetting across all orders, demonstrating the strongest task-order robustness.

Such insensitivity to task order is particularly important for real-world medical continual learning, where data from different hospitals, modalities, and anatomical regions rarely arrive in a fixed or predetermined sequence.

D.3. The Confidence Threshold τ_t for Each Task

Table 9 presents the calibrated thresholds τ_t discussed in the router threshold ablation study. These thresholds serve as the critical decision boundaries for the VAE Router to distinguish between known tasks and out-of-distribution (OOD) samples. As detailed in the implementation settings, these values were derived via 5-fold cross-validation on the training set of each task based on the in-distribution ELBO

Algorithm 7. Continual Alignment for SAM

Input: Pre-trained SAM θ, Tasks $\mathcal{D} = \{D_1^{tr}, \ldots, D_N^{tr}\}$
Output: Alignment Layer Pool $\mathcal{P}_\mathcal{A} = \{\mathcal{A}_t\}_{t=1}^N$,
VAE Router $\mathcal{P}_\mathcal{V} = \{\mathcal{V}_t\}_{t=1}^N$, Thresholds $\mathcal{T} = \{\tau_t\}_{t=1}^N$
Initialize $\mathcal{P}_\mathcal{A} \leftarrow \emptyset$, $\mathcal{P}_\mathcal{V} \leftarrow \emptyset$, $\mathcal{T} \leftarrow \emptyset$.
for $t = 1, \ldots, N$ **do**
  for $(x, y) \in D_t^{tr}$ **do**
    $Z = E_\theta(x)$; $\tilde{Z} = \mathcal{A}_t(Z)$. $\hat{m} = D_\theta(\tilde{Z})$.
    Update $\mathcal{A}_t$ with loss $\mathcal{L}_{align}(\hat{m}, y)$.
  end for
  Train VAE $\mathcal{V}_t$ on D_t^{tr} using Z.
  Compute τ_t for $\mathcal{V}_t$
  Save $\mathcal{A}_t$, $\mathcal{V}_t$, τ_t into pools $\mathcal{P}_\mathcal{A}$, $\mathcal{P}_\mathcal{V}$, $\mathcal{T}$.
end for
return $\mathcal{P}_\mathcal{A}, \mathcal{P}_\mathcal{V}, \mathcal{T}$

Inference Phase: Given a test image I_{test}.
  $Z_{test} = E_\theta(I_{test})$.
  for $t = 1, \ldots, N$ **do**
    Compute $s_t = \mathcal{L}_{VAE_t}(Z_{test})$.
  end for
  $t^* = \arg\min_t s_t$.
  if $s_{t^*} < \tau_{t^*}$
    Load $\mathcal{A}_{t^*}$; $\tilde{Z}_{test} = \mathcal{A}_{t^*}(Z_{test})$.
  else
    Load Identity Alignment Layer$\mathcal{A}^*$
    $\tilde{Z}_{test} = \mathcal{A}^*(Z_{test})$.
  end if
  $\hat{m}_{test} = D_\theta(\tilde{Z}_{test})$.

score distribution. We report the thresholds calculated using four different statistical criteria: $\mu + 2\sigma$, p_{95}, p_{97}, and p_{99}. For CA-SAM, we adopt the p_{97} strategy, which was experimentally determined to offer the optimal trade-off between preserving known-task segmentation performance and effectively rejecting unseen domains.

E. Explainability Metrics Computation

We selected the TV (Total Variation) and JS (Jensen-Shannon) divergence metrics for comparison. The two metrics are computed as follows:

$$D_{TV}(P, Q) = \frac{1}{2} \sum_x |P(x) - Q(x)| \tag{13}$$

$$D_{JS}(P, Q) = \frac{1}{2} \left(D_{KL}(P \parallel M) + D_{KL}(Q \parallel M) \right) \tag{14}$$

where $M = \frac{1}{2}(P + Q)$ is the average of the two distributions, and D_{KL} is the KL divergence.

F. Ablation Study

F.1. Cross-dataset Stability Ablation

Temperature Ablation of Attention Pooling. The temperature parameter T in the attention pooling mechanism

Method	IoU on Med			BIoU on Med		
	Last-IoU ↑	Avg-IoU ↑	FF-IoU ↓	Last-BIoU ↑	Avg-BIoU ↑	FF-BIoU ↓
DN → 56Nx → EBHI-SEG → STS-2D → promise12 → MSD_Spleen → MSD_Prostate → Polyp → ACDC						
SAM [17] + AL(naive)	25.27	34.06	54.13%	25.15	25.94	37.12%
LwF [19]	30.19	45.27	6.55%	18.81	30.08	8.96%
EWC [18]	25.38	31.01	46.68%	19.30	23.00	32.44%
ER [21]	73.73	74.68	6.20%	51.03	53.00	11.50%
L2P [30]	67.65	70.48	2.47%	41.71	43.78	4.68%
MoDA [35]	65.83	67.02	2.20%	49.75	48.67	1.12%
CA-SAM (Ours)	**75.61**	**75.38**	**1.71%**	**58.94**	**57.30**	**1.84%**
EBHI-SEG → Polyp → ACDC → MSD_Prostate → 56Nx → MSD_Spleen → DN → promise12 → STS-2D						
SAM [17] + AL(naive)	13.13	31.98	66.10%	12.01	24.14	49.55%
LwF [19]	25.22	44.26	1.66%	12.99	26.49	2.98%
EWC [18]	18.38	32.81	52.07%	12.29	24.50	38.97%
ER [21]	68.57	68.19	10.92%	47.47	44.99	15.00%
MoDA [35]	67.03	71.97	1.44%	51.44	51.71	1.06%
CA-SAM (Ours)	**75.78**	**76.47**	**1.31%**	**57.73**	**54.79**	**1.64%**
MSD_Prostate → 56Nx → STS-2D → Polyp → DN → MSD_Spleen → ACDC → EBHI-SEG → promise12						
SAM [17] + AL(naive)	28.39	36.85	52.76%	19.54	29.60	45.40%
LwF [19]	13.43	20.58	1.76%	9.84	16.43	1.86%
EWC [18]	22.99	37.88	45.78%	17.83	28.00	27.91%
ER [21]	71.69	62.13	7.76%	53.11	46.74	7.94%
MoDA [35]	67.26	54.03	0.96%	51.54	42.38	0.90%
CA-SAM (Ours)	**76.19**	**63.31**	**1.66%**	**59.51**	**52.74**	**1.70%**

Table 8. The results of different contrastive methods under three different task orders on medical datasets.The best and second best performances are highlighted.

Parameter	ACDC	EBHI-SEG	56Nx	DN	Polyp	MSD-Prostate	MSD-Spleen	Promise12	STS-2D
$\mu + 2\sigma$	0.0790	0.0680	0.2121	0.1589	0.1462	0.1857	0.1257	0.1638	0.0632
p_{95}	0.0761	0.0653	0.2066	0.1553	0.1446	0.1817	0.1199	0.1404	0.0622
p_{97}(Ours)	**0.0813**	**0.0690**	**0.2258**	**0.1646**	**0.1494**	**0.1915**	**0.1285**	**0.1483**	**0.0662**
p_{99}	0.0916	0.0806	0.2808	0.1819	0.1766	0.2062	0.1544	0.1782	0.0729

Table 9. Calibrated Thresholds (τ_t) for Each Dataset under Different Strategies.

controls the smoothness of the Softmax function used for spatial feature aggregation. To evaluate the sensitivity of our framework to this hyperparameter, we conducted an ablation study by varying T from 0.1 to 16, as detailed in Table 10. The experimental results demonstrate that our VAE Router exhibits strong robustness to variations in temperature. Across a wide range of values ($T \in [1, 16]$), the model consistently maintains high task identification performance, with Zero-shot Accuracy stabilizing above 97% and Average IoU remaining steady. Therefore, since the framework exhibits such strong robustness within this broad range, we employ the standard and computationally simple setting of $T = 1$ as the default configuration in our main experiments, prioritizing model simplicity without sacrificing performance.

VAE Structure Ablation. To investigate the role of the KL divergence in our task routing mechanism, we conducted a sensitivity analysis on the regularization coeffi-

Parameter	IoU on Med			BIoU on Med			Zero-shot				
	Last	Avg	FF	Last	Avg	FF	IoU	BIoU	Acc	IoU(Med)	BIoU(Med)
$T = 0.1$	76.04	76.79	1.58%	60.05	59.55	0.27%	55.68	58.15	99.03	76.07	59.99
$T = 0.25$	75.85	76.77	1.62%	59.92	59.52	0.13%	55.99	58.33	99.49	75.93	59.83
$T = 0.5$	75.54	76.55	1.87%	60.04	59.54	0.14%	55.24	57.90	98.36	75.91	60.03
$T = 1$	75.73	76.75	1.78%	59.80	59.51	0.29%	54.99	57.74	97.86	75.62	59.57
$T = 2$	76.05	76.88	1.32%	59.48	59.21	0.38%	55.72	58.14	99.06	76.41	59.79
$T = 4$	76.07	76.97	1.42%	59.74	59.41	0.35%	55.34	57.88	98.42	**76.46**	59.96
$T = 8$	**76.24**	**76.99**	**1.23%**	59.82	59.39	0.20%	**56.00**	**58.34**	**99.53**	76.38	59.76
$T = 16$	75.78	76.75	1.68%	**60.07**	**59.57**	**0.07%**	55.47	58.02	98.68	76.12	**60.19**

Table 10. Ablation Study on Attention Pooling Temperature Parameter T

Parameter	IoU on Med			BIoU on Med			Zero-shot				
	Last	Avg	FF	Last	Avg	FF	IoU	BIoU	Acc	IoU(Med)	BIoU(Med)
$\beta = 0$	44.81	52.55	4.25%	36.74	42.08	3.55%	46.64	53.56	84.59	53.25	44.14
$\beta = 1$	44.64	49.73	5.17%	36.39	40.18	4.51%	36.69	41.11	14.07	44.24	36.60
$\beta = 1.5$	48.35	55.01	6.12%	39.44	44.10	4.73%	36.88	41.23	11.62	52.32	42.37
$\beta = 2$	58.75	62.70	3.43%	47.06	49.49	2.96%	39.66	44.41	29.36	62.12	49.91
$\beta = 2.5$	64.26	67.03	3.15%	51.49	52.69	2.37%	44.42	49.75	56.18	66.18	53.09
$\beta = 3$	68.48	70.46	1.96%	53.75	54.50	1.64%	49.59	53.99	84.14	71.75	56.24
$\beta = 3.5$	68.80	70.17	1.22%	54.19	54.69	1.26%	48.49	52.82	72.24	71.70	56.81
$\beta = 4$	69.52	71.37	1.35%	55.50	55.89	0.97%	51.14	55.52	88.42	72.40	57.56
$\beta = 4.5$	73.13	74.75	0.82%	57.51	57.80	0.61%	52.86	56.74	94.15	74.54	58.54
$\beta = 5$	72.79	74.44	1.33%	57.53	57.71	0.77%	52.68	56.75	94.30	75.18	59.49
$\beta = 5.5$	73.19	74.41	**0.51%**	58.01	57.89	0.40%	53.56	57.20	95.58	74.44	58.78
$\beta = 6$	73.54	74.99	0.80%	57.97	58.21	0.85%	54.13	57.51	96.52	74.98	59.40
$\beta = 6.5$	74.43	75.86	1.72%	58.76	58.85	0.57%	53.74	57.23	95.88	76.08	60.00
$\beta = 7$	73.86	75.25	0.87%	58.46	58.42	0.68%	55.48	58.09	98.63	75.61	59.79
$\beta = 7.5$	74.26	75.46	0.69%	58.52	58.58	0.55%	55.11	57.85	98.00	75.40	59.43
$\beta = 8$	75.25	76.25	1.14%	59.14	58.86	0.39%	52.79	55.62	93.04	75.70	59.46
$\beta = 8.5$	75.72	76.47	0.75%	59.37	58.99	0.30%	55.77	58.23	99.09	**76.34**	59.95
$\beta = 9$	75.06	76.24	1.64%	59.11	59.02	0.51%	55.34	57.96	98.48	75.84	59.67
$\beta = 9.5$	75.04	76.17	1.06%	59.45	59.14	0.32%	55.73	58.18	99.09	75.75	59.77
$\beta = 10$	75.76	76.78	1.60%	59.58	59.24	0.27%	55.55	58.11	98.69	76.25	59.93
$\beta = 10.5$	75.41	76.38	1.53%	59.60	59.34	0.40%	55.73	58.18	99.02	76.02	60.01
$\beta = 11$	75.66	76.64	1.53%	59.72	59.30	0.23%	55.41	58.11	98.70	75.98	59.80
$\beta = 11.5$	75.25	76.33	1.73%	59.61	59.32	0.39%	55.88	58.27	99.32	76.20	**60.07**
$\beta = 12$	75.60	76.61	1.49%	59.46	59.22	0.38%	56.01	58.36	99.53	75.77	59.60
$\beta = 12.5$	75.48	76.81	2.03%	59.41	59.33	0.55%	54.85	57.78	97.84	75.75	59.43
$\beta = 13$	75.42	76.46	1.40%	59.60	59.30	0.27%	54.99	57.88	97.99	76.06	60.03
$\beta = 13.5$	75.73	76.74	1.73%	59.78	59.49	0.32%	55.65	58.11	98.98	76.04	60.03
$\beta = 14$	75.81	76.85	1.56%	59.46	59.25	0.36%	55.95	58.26	99.41	75.92	59.53
$\beta = 14.5$	75.81	76.88	1.66%	59.70	59.45	0.27%	56.01	58.34	99.57	75.99	59.57
$\beta = 15$	75.44	76.58	1.84%	59.70	59.49	0.35%	**56.18**	**58.44**	**99.79**	75.91	59.98
$\beta = 15.5$	75.77	76.69	1.82%	59.94	59.51	0.22%	55.71	58.19	99.11	75.68	59.64
$\beta = 16$	76.05	76.90	1.47%	59.98	**59.56**	0.17%	55.42	57.91	98.66	75.87	59.86
$\beta = 16.5$	**76.12**	**76.90**	1.43%	59.95	59.45	0.24%	55.63	58.13	99.05	76.23	59.89
$\beta = 17$	76.00	76.90	1.37%	59.83	59.50	0.27%	55.98	58.27	99.41	76.14	59.89
$\beta = 17.5$	75.88	76.68	1.71%	**60.03**	59.52	**0.05%**	55.85	58.23	99.30	75.78	59.81
$\beta = 18$	75.58	76.56	1.73%	59.74	59.38	0.21%	55.97	58.33	99.48	75.71	59.71

Table 11. Ablation study on the KL regularization coefficient β in the VAE Router.

cient β in our ELBO loss, varying it from 0 to 18. As shown in Table 11, β is pivotal in balancing feature reconstruction and the constraint on the latent space distribution. In the lower range ($\beta < 6$), the model exhibits significant instability. As β increases, the stronger KL penalty drives the formation of more separated and structured distribution boundaries for each task in the latent space. This consequently enforces a distinct and compact feature distribution for every task, which significantly enhances the discriminability between different tasks and is crucial for the router to reject OOD samples. We observe a substantial performance stabilization for $\beta \in [7, 18]$, where the Average IoU on Med consistently stays above 75% and Zero-shot Accuracy exceeds 98%.