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.

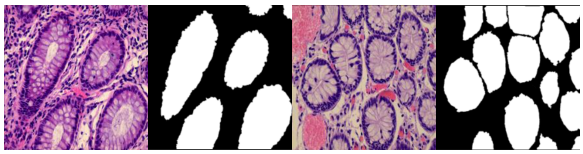
MSD_Spleen: The MSD_Spleen dataset contains 876 training images and 146 test images. The dataset modality is CT imaging, and the segmentation target is the spleen.

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.

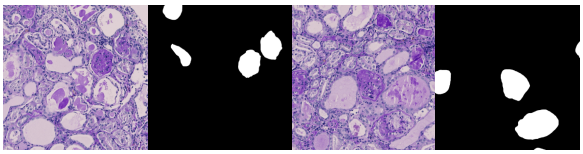
STS-2D: The STS-2D dataset consists of 1700 training images and 70 test images. The dataset modality is X-ray imaging, and the segmentation target is the teeth.

Figure 1 present representative examples of the images and corresponding masks for each dataset.

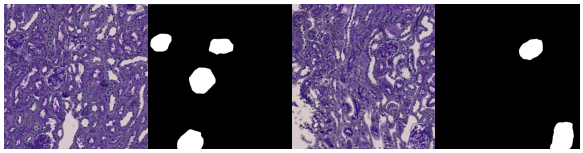
(1) ACDC



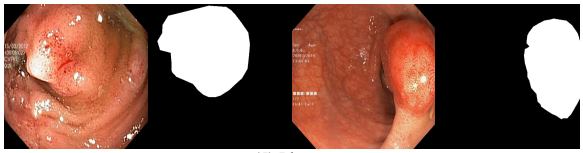
(2) EBHI-SEG



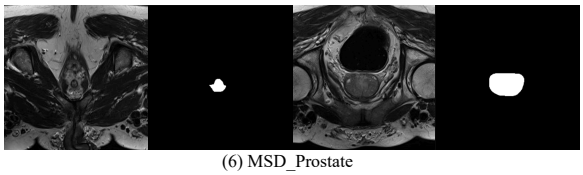
(3) 56Nx



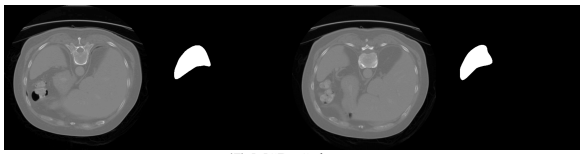
(4) DN



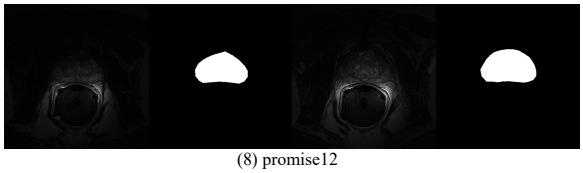
(5) Ployp



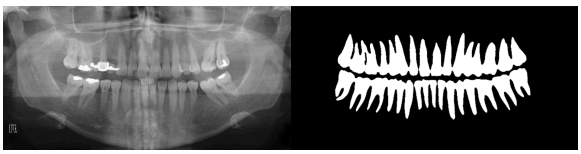
(6) MSD_Prostate



(7) MSD_Spleen



(8) promise12



(9) STS-2D

Figure 1. **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,

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

where $\mathrm{IoU}_{k,t}$ and $\mathrm{BIoU}_{k,t}$ 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

$$\text{Last-IoU} = \mathrm{IoU}_N, \quad \text{Last-BIoU} = \mathrm{BIoU}_N. \tag{2}$$

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:

$$\text{Avg-IoU} = \frac{1}{N}\sum_{t=1}^{N} \mathrm{IoU}_t, \quad \text{Avg-BIoU} = \frac{1}{N}\sum_{t=1}^{N} \mathrm{BIoU}_t. \tag{3}$$

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

$$f_{k,t} = \max_{j\in\{1,\ldots,t-1\}} \left(\mathrm{IoU}_{k,j} - \mathrm{IoU}_{k,t}\right). \tag{4}$$

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

$$\text{FF-IoU} = \frac{1}{N-1}\sum_{k=1}^{N-1} f_{k,N}. \tag{5}$$

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

C. Single-dataset Versatility Analysis

Table 1 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 [3]	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 [2]	5.14M	77.05	85.07	47.11	58.34	78.46	**72.33**	90.27	87.65	81.88	72.91
SAMMed 2D [1]	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 [3]	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 [2]	5.14M	73.18	50.18	33.46	43.04	56.06	**65.24**	86.52	71.97	79.63	58.41
SAMMed 2D [1]	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 1. 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 2, 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 3 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:

$$D_{TV}(P, Q) = \frac{1}{2} \sum_x |P(x) - Q(x)| \tag{6}$$

$$D_{JS}(P, Q) = \frac{1}{2} \left(D_{KL}(P \parallel M) + D_{KL}(Q \parallel M) \right) \tag{7}$$

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 [3] + AL(naive)	25.27	34.06	54.13%	25.15	25.94	37.12%
LwF [5]	30.19	45.27	6.55%	18.81	30.08	8.96%
EWC [4]	25.38	31.01	46.68%	19.30	23.00	32.44%
ER [6]	73.73	74.68	6.20%	51.03	53.00	11.50%
L2P [7]	67.65	70.48	2.47%	41.71	43.78	4.68%
MoDA [8]	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 [3] + AL(naive)	13.13	31.98	66.10%	12.01	24.14	49.55%
LwF [5]	25.22	44.26	1.66%	12.99	26.49	2.98%
EWC [4]	18.38	32.81	52.07%	12.29	24.50	38.97%
ER [6]	68.57	68.19	10.92%	47.47	44.99	15.00%
MoDA [8]	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 [3] + AL(naive)	28.39	36.85	52.76%	19.54	29.60	45.40%
LwF [5]	13.43	20.58	1.76%	9.84	16.43	1.86%
EWC [4]	22.99	37.88	45.78%	17.83	28.00	27.91%
ER [6]	71.69	62.13	7.76%	53.11	46.74	7.94%
MoDA [8]	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 2. 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 3. 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 4. 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 4. 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 5. 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 5, 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%.

References

[1] 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. 3

[2] 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. 3

[3] 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. 3, 5

[4] 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. 5

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

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

[7] 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. 5

[8] 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. 5