Mamba Goes HoME: Hierarchical Soft Mixture-of-Experts for 3D Medical Image Segmentation

Szymon Płotka[1,2]* **Gizem Mert**[3] **Maciej Chrabaszcz**[4,5]
Ewa Szczurek[1,3] **Arkadiusz Sitek**[6,7]

[1] Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Poland
[2] Faculty of Mathematics and Computer Science, Jagiellonian University, Poland
[3] Institute of AI for Health, Helmholtz Munich, Germany
[4] Faculty of Electronics and Information Technology, Warsaw University of Technology, Poland
[5] NASK - National Research Institute, Poland
[6] Faculty of Radiology, Massachusetts General Hospital, USA
[7] Department of Radiology, Harvard Medical School, USA

Abstract

In recent years, artificial intelligence has significantly advanced medical image segmentation. Nonetheless, challenges remain, including efficient 3D medical image processing across diverse modalities and handling data variability. In this work, we introduce Hierarchical Soft Mixture-of-Experts (HoME), a two-level token-routing layer for efficient long-context modeling, specifically designed for 3D medical image segmentation. Built on the Mamba Selective State Space Model (SSM) backbone, HoME enhances sequential modeling through adaptive expert routing. In the first level, a Soft Mixture-of-Experts (SMoE) layer partitions input sequences into local groups, routing tokens to specialized per-group experts for localized feature extraction. The second level aggregates these outputs through a global SMoE layer, enabling cross-group information fusion and global context refinement. This hierarchical design, combining local expert routing with global expert refinement, enhances generalizability and segmentation performance, surpassing state-of-the-art results across datasets from the three most widely used 3D medical imaging modalities and varying data qualities. The code is publicly available at github.com/gmum/MambaHoME.

1 Introduction

Three-dimensional (3D) medical image segmentation lies at the core of computer-aided diagnosis, image-guided interventions, and treatment planning across modalities such as Computed Tomography (CT) [51, 12, 44], Magnetic Resonance Imaging (MRI) [33, 48, 21], and Ultrasound (US) [60, 24, 6]. A key characteristic of medical imaging data is its hierarchical structure: local patterns, such as tumor lesions, are embedded within larger anatomical structures like organs, which themselves follow a consistent global arrangement [20]. We hypothesize that *models capable of capturing these local-to-global spatial hierarchies in 3D medical data can enhance segmentation performance and yield latent representations that generalize effectively across diverse imaging modalities*.

Convolutional Neural Networks (CNNs) provide local feature extraction with linear complexity in the number of input pixels, but their limited receptive fields hinder their ability to capture global

*Correspondence to: Szymon Płotka (SZYMON.PLOTKA@UJ.EDU.PL)

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

spatial patterns [42, 57, 22, 52, 45]. In contrast, Vision Transformers (ViTs) [16, 5, 31] leverage global attention mechanisms to model long-range dependencies. However, their quadratic complexity in the number of tokens makes them computationally costly for high-resolution 3D data, posing significant challenges for scalability. Moreover, while many ViT-based models [53, 18, 38, 30, 27, 47] incorporate multi-scale feature extraction and attention mechanisms, they still struggle to effectively aggregate fine-to-coarse semantic information, particularly in dense prediction tasks such as volumetric segmentation. Although these models demonstrate improved performance, they do not explicitly model global and local spatial patterns and their mutual arrangement. Modeling the transition from local to global patterns requires processing long sequences and capturing long-range dependencies, imposing prohibitive memory and computational demands on current architectures [9].

Recently, Selective State Space Models (SSMs), such as Mamba [19, 15], have emerged as efficient alternatives to ViTs by offering linear complexity in the number of tokens to capture long-range dependencies, including in 3D medical imaging [54, 46, 29, 11, 50]. Although SSMs effectively capture global context with lower computational cost than attention-based methods, they are not inherently designed to adaptively handle diverse local patterns in medical data. Efficient management of such local patterns in complex, multi-scale data while maintaining scalability is achieved through Mixture-of-Experts (MoE) frameworks, which dynamically route features to specialized subnetworks. MoE-based methods have gained prominence across domains such as language modeling [4, 56, 26], vision tasks [39, 59, 41, 14, 17], multimodal learning [34, 8, 55], and medical applications [25, 13, 49, 38]. However, combining the global efficiency of SSMs with the localized adaptability of MoE remains largely unexplored, particularly in 3D medical imaging, where balancing efficiency and generalization across multi-modal datasets under resource constraints is paramount.

To address these challenges, we introduce the first-in-class model that smoothly integrates Mamba with hierarchical Soft MoE in a multi-stage network for local-to-global pattern modeling and 3D image segmentation with competitive memory and compute efficiency, while maintaining state-of-the-art segmentation performance. Specifically, our contributions are as follows:

1. We introduce **H**ierarchical **So**ft **M**ixture-of-**E**xperts (**HoME**), a two-level, token-routing MoE layer for efficient capture of local-to-global pattern hierarchies, where tokens are grouped and routed to local experts in the first SMoE level, then aggregated and passed via a global SMoE in the second level,
2. We design a unified architectural block that integrates Mamba's SSMs with HoME, combining memory-efficient long-sequence processing with hierarchical expert routing,
3. We embed the above novel solutions into a multi-stage U-shaped architecture, called **Mamba-HoME**, specifically designed for 3D medical image segmentation, where the integration of hierarchical memory and selective state space modeling enhances contextual representation across spatial and depth dimensions.

Through comprehensive experiments on four publicly available datasets, including PANORAMA [2], AMOS [23], FeTA 2022 [37], and MVSeg [10] as well as one in-house CT dataset, we demonstrate that Mamba-HoME outperforms current state-of-the-art methods in both segmentation accuracy and computational efficiency, while generalizing effectively across three major 3D medical imaging modalities: CT, MRI, and US.

2 Methodology

2.1 Preliminaries

Selective State-Space Models (Mamba). Our network builds on the Mamba layer [19], designed for long-sequence modeling with linear computational complexity. Unlike Transformer-based architectures, which exhibit quadratic complexity with respect to sequence length, Mamba achieves linear-time processing through a continuous-time recurrent formulation with input-dependent parameters. Given a feature sequence $z = \{z_t\}_{t=1}^{N} \in \mathbb{R}^{B\times N\times d}$, the hidden state h_t and output y_t are updated as follows:

$$h_t = A(z_t)h_{t-1} + B(z_t)z_t, \qquad y_t = C(z_t)h_t, \tag{1}$$

where $A(\cdot), B(\cdot), C(\cdot)$ are learnable input-dependent linear mappings. This formulation allows Mamba to capture both short- and long-range dependencies with a memory and computational

complexity of $\mathcal{O}(Nd)$. In our architecture, the Mamba layer serves as a backbone for volumetric sequence modeling, efficiently capturing spatial dependencies across 3D data inputs.

Soft Mixture-of-Experts (SMoE). SMoE framework [39] introduces a modular computation strategy where each input token is dynamically routed to multiple experts, and their outputs are combined using soft routing weights. Given an input tensor $x \in \mathbb{R}^{B\times N\times d}$ and E experts, a gating network computes routing probabilities $p_{b,n,e}$ for each expert e and token $x_{b,n}$:

$$y_{b,n} = \sum_{e=1}^{E} p_{b,n,e} \cdot \mathrm{FFN}_e(x_{b,n}), \qquad y_{b,n} \in \mathbb{R}^d, \tag{2}$$

where FFN_e denotes the e-th expert network and $\sum_{e=1}^{E} p_{b,n,e} = 1$ ensures a valid soft assignment. Each expert is typically a small feed-forward network, and the gating network is a learned per-token function (e.g., an MLP) that assigns weights dynamically.

While SMoE improves model capacity and modularity, global routing over all tokens incurs a computational cost of $\mathcal{O}(NdE)$, which becomes prohibitive for high-resolution 3D inputs where N can reach millions. Moreover, SMoE lacks spatial locality and hierarchical aggregation, both of which are important for structured volumetric data.

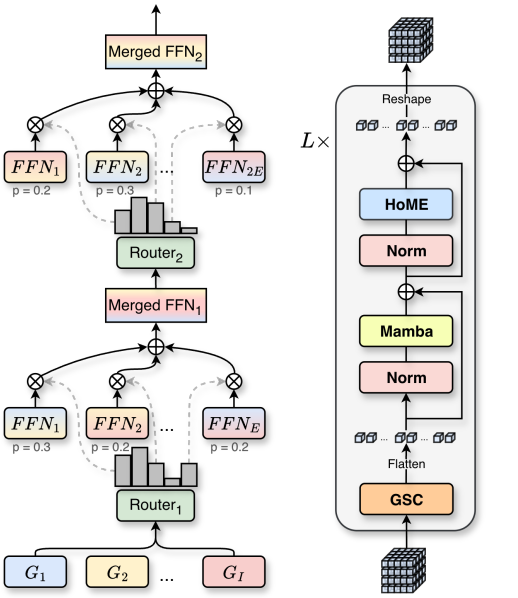

Figure 1: An overview of the HoME layer and Mamba-HoME Block design. (a) The HoME layer operates on G_I groups of K_I tokens. Router_1 routes each group to E local experts for intra-group feature extraction, producing aggregated slot representations. Router_2 routes these aggregated slots to $2E$ experts for inter-group communication and global refinement. (b) The Mamba-HoME Block combines a Gated Spatial Convolution (GSC) module, Mamba for efficient long-sequence processing, and hierarchical expert processing (HoME). Dynamic Tanh is used for normalization to improve gradient stability and efficiency.

Notation. At encoder stage $i \in \{1, \ldots, I\}$, the input feature sequence is $x \in \mathbb{R}^{B\times N_i\times d}$, where B is the batch size, N_i is the sequence length, and d is the feature dimension. The sequence is divided into $G_i = \lceil N_i/K_i \rceil$ groups of up to K_i tokens each. Each group is processed by E_i local experts, each maintaining S_i learnable slots. The total number of expert-slot pairs is $M_i = E_i \cdot S_i$, and the slot embeddings are $E_{\text{slots}}^{(i)} \in \mathbb{R}^{E_i\times S_i\times d}$. Tokens are assigned to slots using weights $A_{b,g,k,m}$. The first-level outputs of local experts are $y^{(1)}$, which are then refined by a larger set of second-level experts $E_{2,i} = 2E_i$ to produce the final outputs $y^{(2)}$.

2.2 Hierarchical Soft Mixture-of-Experts (HoME)

We introduce the HoME layer (see Figure 1(a)), which enhances feature processing through a hierarchical two-level structure. HoME extends the SMoE concept with a hierarchical two-level routing structure that processes grouped tokens locally and enables inter-group information exchange efficiently. It comprises three key steps: (1) grouped slot assignment for token processing, (2) first-level MoE processing for local feature extraction, and (3) second-level MoE refinement for global feature extraction, allowing inter-group communication.

Grouped Slot Assignment. In hierarchical vision encoders, particularly for dense 3D inputs (e.g., volumetric data), the sequence length is highest in early stages due to large spatial resolutions. Global expert routing on these long sequences causes high computational and memory costs, as each token is compared to all expert slots, yielding a complexity of $\mathcal{O}(N_i \cdot M_i)$, where N_i is the sequence length at stage i and M_i is the total number of expert slots. To address this scalability bottleneck, we introduce Grouped Slot Assignment, a locality-aware routing mechanism that divides the input

sequence into groups and performs soft assignment independently per group. We define the group size as $K_i = K_1 \cdot \rho^{i-1}$ $(0 < \rho < 1)$, and zero-pad sequences to $N_i' = G_i \cdot K_i$, giving $\hat{x} \in \mathbb{R}^{B \times N_i' \times d}$.

Each group $x_g \in \mathbb{R}^{B \times K_i \times d}$ is routed independently. Assignment logits are computed via a dot product between each token and learnable expert-specific slot embeddings $E_{\text{slots}}^{(i)} \in \mathbb{R}^{E_i \times S_i \times d}$.

Let $e \in \{1, \ldots, E_i\}$ and $s \in \{1, \ldots, S_i\}$. The logits are as follows:

$$S_{b,g,k,e,s} = \sum_{j=1}^{d} x_{b,g,k,j} \cdot E_{e,s,j}^{(i)}, \quad S \in \mathbb{R}^{B \times G_i \times K_i \times E_i \times S_i}. \tag{3}$$

An optional binary mask $M \in \{0,1\}^{B \times N_i}$, indicating valid (unpadded) tokens, is extended to $\hat{M} \in \{0,1\}^{B \times N_i'}$ after padding. To prevent padded tokens from affecting expert assignment, the logits for invalid tokens are masked by setting:

$$S_{b,g,k,e,s} \;=\; \begin{cases} S_{b,g,k,e,s}, & \text{if } \hat{M}_{b,\, gK_i+k} = 1, \\ -\infty, & \text{otherwise.} \end{cases} \tag{4}$$

Expert-slot pairs are flattened ($m = (e-1)S_i + s$, $M_i = E_i \cdot S_i$), and normalized with softmax:

$$A_{b,g,k,m} = \frac{\exp(S_{b,g,k,m})}{\sum_{m'=1}^{M_i} \exp(S_{b,g,k,m'})}, \qquad A \in \mathbb{R}^{B \times G_i \times K_i \times M_i}. \tag{5}$$

Each slot representation is computed as a weighted aggregation of tokens within its group:

$$\tilde{x}_{b,g,e,s,j} = \sum_{k=1}^{K_i} A_{b,g,k,m} \cdot x_{b,g,k,j}, \qquad \tilde{x} \in \mathbb{R}^{B \times G_i \times E_i \times S_i \times d}. \tag{6}$$

By performing routing within groups, peak memory usage is reduced and locality is preserved, enabling efficient hierarchical token-to-expert assignment. While the total computational complexity remains $\mathcal{O}(N_i M_i d)$, the smaller group-wise computations allow scalable processing of long sequences while being memory-efficient.

Hierarchical Expert Processing. Let $\tilde{x}^\flat \in \mathbb{R}^{B \times G_i \times M_i \times d}$ denote the grouped slot representations after flattening the (e, s) dimensions of $\tilde{x}$. The first level routes slots within each group to a subset of E_i experts, promoting local specialization and feature refinement. This produces group-specific outputs while preserving slot structure. For each group $g \in \{1, \ldots, G_i\}$, the gating network computes routing weights for the E_i experts $(\text{FFN}_1, \text{FFN}_2, \ldots, \text{FFN}_{E_i})$:

$$\text{Router}_1(\tilde{x}_{b,g}^\flat; \theta_{\text{gate}}^{(1)}) \;=\; \text{softmax}\left(\text{MLP}\left(\frac{1}{M_i}\sum_{m=1}^{M_i} \tilde{x}_{b,g,m}^\flat\right)\right), \tag{7}$$

where $\tilde{x}_{b,g}^\flat = \{\tilde{x}_{b,g,m}^\flat\}_{m=1}^{M_i} \in \mathbb{R}^{M_i \times d}$, $\theta_{\text{gate}}^{(1)}$ are gating parameters, and the softmax normalizes over the expert dimension, yielding $\text{Router}_1 \in \mathbb{R}^{E_i}$. Each expert, implemented as an $\text{FFN}_e : \mathbb{R}^{M_i \times d} \to \mathbb{R}^{M_i \times d}$, processes the group's slots independently. The output for expert $e \in \{1, \ldots, E_i\}$ is

$$y_{b,g,e} \;=\; \left[\text{Router}_1(\tilde{x}_{b,g}^\flat; \theta_{\text{gate}}^{(1)})\right]_e \cdot \text{FFN}_e(\tilde{x}_{b,g}^\flat), \tag{8}$$

and the aggregated output is:

$$y_{b,g}^{(1)} \;=\; \sum_{e=1}^{E_i} y_{b,g,e}, \qquad y^{(1)} \in \mathbb{R}^{B \times G_i \times M_i \times d}.$$

Dynamic Slot Refinement. The second level routes group slots to $E_{2,i}$ experts for global feature refinement. The tensor $y^{(1)}$ is transformed into $\tilde{y} \in \mathbb{R}^{B \times (G_i M_i) \times d}$ before being passed through the second-level experts. The gating network computes routing weights for $E_{2,i}$ experts $(\text{FFN}_1, \text{FFN}_2, \ldots, \text{FFN}_{E_{2,i}})$:

$$\text{Router}_2(\tilde{y}; \theta_{\text{gate}}^{(2)}) \;=\; \text{softmax}\big(\text{MLP}(\tilde{y})\big), \tag{9}$$

where $\theta^{(2)}_{\text{gate}}$ are the gating parameters, producing $\text{Router}_2 \in \mathbb{R}^{B\times(G_iM_i)\times E_{2,i}}$. Each second-level expert, implemented as $\text{FFN}_{e_2} : \mathbb{R}^{(G_iM_i)\times d} \to \mathbb{R}^{(G_iM_i)\times d}$, processes $\tilde{y}$. The output for expert $e_2 \in \{1, \ldots, E_{2,i}\}$ is:

$$y_{b,e_2} = \left[\text{Router}_2(\tilde{y}_b; \theta^{(2)}_{\text{gate}})\right]_{e_2} \cdot \text{FFN}_{e_2}(\tilde{y}_b), \tag{10}$$

and outputs are aggregated into:

$$y^{(2)} = \sum_{e_2=1}^{E_{2,i}} y_{b,e_2}, \qquad y^{(2)} \in \mathbb{R}^{B\times(G_iM_i)\times d}.$$

The final stage reconstructs a structured output using an attention-based combination to emphasize relevant slots and remove padding, yielding a compact, task-aligned representation. The tensor $y^{(2)}$ is reshaped to $\mathbb{R}^{B\times G_i\times M_i\times d}$. For batch element b, group g, and token $k \in \{1, \ldots, K_i\}$, the output is

$$y_{b,g,k} = \sum_{m=1}^{M_i} W_{b,g,k,m} \cdot y^{(2)}_{b,g,m}, \tag{11}$$

where $y^{(2)}_{b,g,m} \in \mathbb{R}^d$, and $W_{b,g,k,m} \in [0, 1]$ are attention weights satisfying $\sum_{m=1}^{M_i} W_{b,g,k,m} = 1$. After undoing padding, the final output is $\hat{y} \in \mathbb{R}^{B\times N_i\times d}$.

2.3 Mamba-HoME Block

Figure 1(b) provides an overview of the proposed Mamba-HoME Block (Mamba-HoMEB). The Mamba-HoMEB extends the Mamba layer by incorporating hierarchical downsampling, Gated Spatial Convolution (GSC) module [54], and a Hierarchical Soft Mixture-of-Experts (HoME) layer (see Section 2.2). To improve gradient stability and computational efficiency in SSMs, we use Dynamic Tanh (DyT) [58] normalization, defined as:

$$f_{\text{DyT}}(x) = w \cdot \tanh(\alpha \cdot x) + b, \tag{12}$$

where $x \in \mathbb{R}^{B\times d}$ is the input tensor, $w, b \in \mathbb{R}^d$ are learnable per-channel vector parameters, and $\alpha \in \mathbb{R}$ is a shared scalar. DyT applies a point-wise nonlinearity, leveraging the bounded nature of $\tanh$ to stabilize gradients. Unlike Layer Normalization [7], it avoids costly mean and variance computations, reducing overhead while maintaining $\mathcal{O}(Bd)$ complexity. DyT accelerates training and validation with performance on par with normalization-based alternatives.

Given a 3D input volume $x \in \mathbb{R}^{B\times C\times D\times H\times W}$, the initial feature map $x^0_1 \in \mathbb{R}^{B\times 48\times \frac{D}{2}\times \frac{H}{2}\times \frac{W}{2}}$ is extracted by a stem layer. This feature map x^0_1 is then passed through each Mamba-HoMEB and its corresponding downsampling layers. For the l-th layer ($l \in \{0, 1, \ldots, L_i - 1\}$) within stage i, the representations for the i-th Mamba-HoMEB are given by:

$$x'^{l}_i = f_{\text{GSC}}(x^l_i), \qquad \bar{x}^l_i = f_{\text{Mamba}}(f_{\text{Norm}}(x'^{l}_i)) + x'^{l}_i, \qquad x^{l+1}_i = f_{\text{HoME}}(f_{\text{Norm}}(\bar{x}^l_i)) + \bar{x}^l_i. \tag{13}$$

where f_{GSC} denotes Gated Spatial Convolution module, f_{Mamba} the Mamba layer, f_{Norm} corresponds to DyT normalization (see Eq. 12), and f_{HoME} the HoME layer. After applying f_{GSC}, the feature map is flattened along the spatial dimensions into a sequence (length N_i) before being processed by f_{Mamba}. The output of the HoME layer is then reshaped back into the original volumetric form to yield x^{l+1}_i.

The HoME layer operates hierarchically at each encoder stage $i \in \{1, \ldots, I\}$, with the number of first-level experts denoted by E_i and the group size (i.e., the number of tokens processed jointly within each local group) denoted by K_i. The number of experts E_i increases monotonically with stage depth, reflecting increased specialization ($E_1 < E_2 < \cdots < E_I$), while the group size K_i decreases, enabling progressively finer-grained processing ($K_1 > K_2 > \cdots > K_I$). In addition to the first-level experts, each stage employs a second-level expert set $E_{2,i}$, which scales proportionally with E_i, i.e., $E_{2,i} = 2E_i$. This second level facilitates global context integration across groups, enhancing inter-group communication. The HoME operation at each stage thus combines local expert routing with global feature aggregation (see Section 2.2).

2.4 The Mamba-HoME Architecture

Building upon SegMamba [54], we introduce a U-shaped encoder-decoder network, called **Mamba-HoME**, designed for 3D medical image segmentation. It leverages a Mamba-based encoder backbone to efficiently capture both long-range dependencies and local features. The model processes an input 3D volume $x \in \mathbb{R}^{B\times C\times D\times H\times W}$ and produces a segmentation mask $y \in \mathbb{R}^{B\times C'\times D\times H\times W}$, where B is the batch size, C is the number of input channels, C' is the number of output classes, and D, H, W are spatial dimensions.

The encoder begins with a stem layer that produces the initial feature map x_1^0 (see Section 2.3), followed by I hierarchical stages. Each encoder stage $i \in \{1, \dots, I\}$ applies a Mamba-HoMEB, producing intermediate representations x_i^l for $l = 0, \dots, L_i - 1$, where L_i denotes the number of layers in the i-th Mamba-HoMEB, and concludes with a downsampling operation:

$$x_{i+1}^0 \;=\; \mathrm{Downsample}_i(x_i^{L_i}), \tag{14}$$

where downsampling halves the spatial dimensions and doubles the channel depth: $D_{i+1} = \lfloor D_i/2 \rfloor$, $H_{i+1} = \lfloor H_i/2 \rfloor$, $W_{i+1} = \lfloor W_i/2 \rfloor$, and $C_{i+1} = 2C_i$. This process results in a sequence of encoder feature maps $\{x_1^{L_1}, x_2^{L_2}, \dots, x_I^{L_I}\}$, with $x_I^{L_I}$ serving as the bottleneck representation.

The decoder mirrors the encoder with $I - 1$ upsampling stages. From the bottleneck $u_0 = x_I^{L_I}$, each decoder stage $j = 1, \dots, I - 1$ fuses the corresponding encoder skip connection with the upsampled decoder features:

$$u_j \;=\; \mathrm{UpBlock}_j\big(x_{I-j}^{L_{I-j}} \oplus \mathrm{Up}(u_{j-1})\big), \tag{15}$$

where $x_{I-j}^{L_{I-j}} \in \mathbb{R}^{B\times C_{I-j}\times D_{I-j}\times H_{I-j}\times W_{I-j}}$ is the skip connection from the encoder stage $I - j$, and $\oplus$ denotes channel-wise concatenation. The upsampled spatial dimensions are $D_j = \lfloor D/2^{I-j} \rfloor$, $H_j = \lfloor H/2^{I-j} \rfloor$, and $W_j = \lfloor W/2^{I-j} \rfloor$. Each $\mathrm{UpBlock}_j$ refines the combined features and maps them to the target resolution for the next decoding stage, and a final prediction layer converts u_{I-1} to the segmentation mask $y \in \mathbb{R}^{B\times C'\times D\times H\times W}$.

The computational complexity of Mamba-HoME scales as $\mathcal{O}(BN_id)$ for the Mamba layer and $\mathcal{O}\big(B\,G_i\,(E_i + E_{2,i})\,L_i\,d\big)$ for the HoME layer at stage i, where N_i is the token sequence length, d is the hidden dimension, G_i is the number of groups, and $E_i + E_{2,i}$ reflects the total number of experts. Compared to Transformer-based models ($\mathcal{O}(BN_i^2d)$), this linear scaling in N_i ensures efficiency for large 3D volumes (e.g., $N_i \approx 10^6$ tokens), with HoME adding a modest overhead proportional to the expert count.

3 Experiments

In this section, we evaluate the performance of the proposed Mamba-HoME against state-of-the-art methods for 3D medical image segmentation (see Section 3.3). We perform an ablation study to understand the importance of different configurations and parameters of Mamba-HoME. Moreover, we compare the segmentation performance of Mamba-HoME trained from scratch with a version pre-trained using a supervised learning approach (see Section 3.4). Finally, in Section 3.5, we investigate how Mamba-HoME generalizes to new modalities. The best results are **bolded**, while the second-best results are underlined.

Table 1: Segmentation performance on the PANORAMA and in-house test sets. PDAC and P denote pancreatic ductal adenocarcinoma and pancreas, respectively. (*) indicates a pre-trained model.

Method	DSC (%) ↑ PDAC	DSC (%) ↑ P	mDSC (%) ↑	mHD95 (mm) ↓	Params (M) ↓	GPU (G) ↓	IS(‡) ↓
VoCo-B (*) [53]	40.5	86.3	71.6	89.2	53.0	17.1	1.5
Hermes [18]	48.2	87.8	75.1	9.3	44.5	17.4	1.9
Swin SMT [38]	49.4	87.0	75.0	32.9	170.8	15.8	1.3
VSmTrans [30]	50.3	87.2	75.4	33.6	47.6	11.1	1.7
uC 3DU-Net [22]	52.0	88.2	76.6	8.1	**21.7**	13.6	**1.0**
Swin UNETR [47]	46.3	87.4	74.2	74.3	72.8	17.1	1.5
SegMamba [54]	49.7	**88.5**	76.0	14.1	66.8	**10.1**	1.2
SuPreM (*) [28]	51.7	88.3	76.6	4.4	62.2	17.1	1.5
Mamba-HoME	54.8	88.3	77.5	6.9	170.1	11.1	1.5
Mamba-HoME (*)	**56.7**	**88.5**	**78.2**	**4.1**	170.1	11.1	1.5

‡ Inference speed (IS) is standardized to 1.0 = 1770 ms.

Table 2: Segmentation performance and generalizability of the Mamba-HoME and previous models on the AMOS dataset (CT and MRI). CT: trained solely on the AMOS-CT subset; CT→MRI: pre-trained on the AMOS-CT subset and fine-tuned on the AMOS-MRI subset; MRI: trained solely on the AMOS-MRI subset; CT+MRI: trained on both subsets. (*) indicates a pre-trained model.

Method	mDSC (%) ↑				mHD95 (mm) ↓			
	CT	CT → MRI	MRI	CT + MRI	CT	CT → MRI	MRI	CT + MRI
Hermes [18]	85.3	84.8	80.7	82.9	40.5	**11.9**	**13.9**	36.1
Swin SMT [38]	85.7	83.2	63.2	81.2	91.4	21.0	45.1	98.9
VSmTrans [30]	85.3	84.0	74.6	78.0	178.7	30.1	30.6	102.1
uC 3DU-Net [22]	82.7	84.5	68.6	84.1	47.3	15.4	16.3	34.9
Swin UNETR [47]	84.3	84.2	75.0	81.2	94.7	28.4	30.7	84.7
SegMamba [54]	86.0	84.4	80.2	84.7	98.9	14.7	33.1	64.2
SuPreM†(*) [28]	86.0	83.9	70.3	83.5	66.0	14.8	52.1	48.2
Mamba-HoME	86.3	84.8	81.0	85.1	45.0	16.0	32.5	54.8
Mamba-HoME (*)	**87.3**	**85.0**	**82.3**	**86.4**	**32.0**	13.3	19.5	**27.7**

† SuPreM was pre-trained on the AMOS-CT training set, while VoCo-B was pre-trained on both the training and validation sets. Here, we removed VoCo-B from the benchmark as it may lead to an unfair comparison.

3.1 Datasets

Pre-training. We use publicly available datasets covering two imaging modalities, including AbdomenAtlas 1.1 [40, 27] for CT scans and TotalSegmentator [1] for MRI scans, both containing voxel-wise annotated masks of abdominal anatomical structures.

Training and fine-tuning. We use datasets covering three primary 3D medical imaging modalities, including publicly available PANORAMA (CT) [2], AMOS (CT and MRI) [23], FeTA 2022 (MRI, fetal) [37], MVSeg (3D US) [10], and an in-house CT dataset. A detailed description of the datasets can be found in the Appendix A.

3.2 Implementation details

The experiments were conducted on a workstation equipped with 8 × NVIDIA H100 GPUs. For implementation, we employ Python 3.11, PyTorch 2.4 [35], and MONAI 1.3.0 within a Distributed Data-Parallel (DDP) training setup.

All training is performed using the $\mathcal{L}_{\text{DiceCE}}$ loss function and the AdamW optimizer, with an initial learning rate of 1e-4 controlled by a cosine annealing scheduler [32], a weight decay of 1e-5, and a batch size of 2. All models are trained with 32-bit floating-point precision to ensure numerical stability and to standardize the training process across all experiments. A detailed description of the implementation can be found in the Appendix B.

3.3 Comparison with state-of-the-art methods

We compare our proposed Mamba-HoME with eight state-of-the-art approaches for 3D medical image segmentation, including uC 3DU-Net [22], Swin SMT [38], VoCo-B [53], SuPreM [28], Hermes [18], Swin UNETR [47], VSmTrans [30], and SegMamba (baseline) [54], across four publicly available and one in-house dataset, covering diverse anatomical structures and imaging modalities, such as CT, MRI, and 3D US. Notably, both VoCo-B and SuPreM are pre-trained on large-scale CT scans using self-supervised and supervised learning approaches, respectively. Additionally, we evaluate Mamba-HoME trained from scratch against Mamba-HoME pre-trained with a supervised learning approach to assess the impact of pretraining on segmentation performance. Detailed quantitative and more qualitative results for benchmarking datasets can be found in the Appendix C, and the Appendix E, respectively.

Table 3: 5-fold cross-validation of segmentation performance on the FeTA 2022 dataset. (*) indicates a pre-trained model.

Method	mDSC (%) ↑	mHD95 (mm) ↓
VoCo-B (*) [53]	86.0	4.0
Hermes [47]	86.5	4.0
Swin SMT [38]	85.9	2.4
VSmTrans [30]	86.1	2.3
uC 3DU-Net [22]	85.9	3.5
Swin UNETR [47]	86.2	2.5
SegMamba [54]	85.9	3.5
SuPreM (*) [28]	85.3	3.6
Mamba-HoME	87.5	2.1
Mamba-HoME (*)	**87.7**	**2.0**

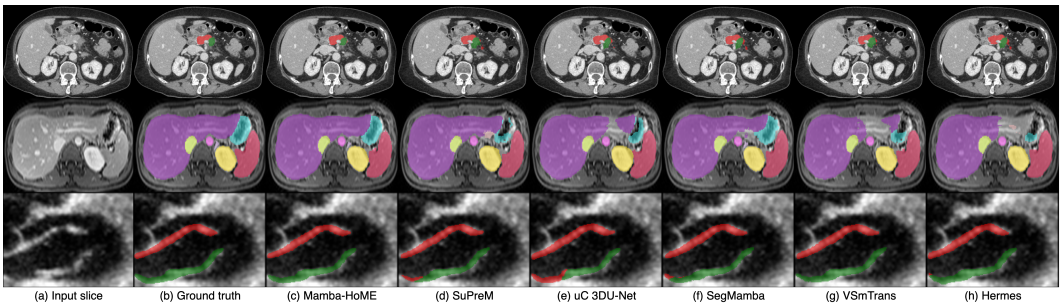

Figure 2: Qualitative segmentation results from top to bottom: CT, MRI, and 3D US. From left to right, each column shows the input slice, ground truth, the proposed Mamba-HoME, and the five next best-performing methods.

Quantitative results. Results for our proposed method, Mamba-HoME, on the PANORAMA and in-house datasets are shown in Table 1. Results for other modalities, including AMOS (CT/MRI), FeTA 2022 (fetal MRI), and MVSeg (3D US), are presented in Table 2, Table 3, and Table 4, respectively.

Our proposed method, Mamba-HoME, demonstrates consistent performance improvements over state-of-the-art baselines across all benchmark datasets and three imaging modalities. Evaluated under two distinct configurations (scratch and pre-trained), Mamba-HoME achieves superior segmentation accuracy, obtaining the best results in terms of both DSC and HD95. Despite having a relatively large number of parameters (170.1M) compared to competing methods, it exhibits low GPU memory usage during inference (see Table 1), a crucial advantage for processing high-resolution 3D medical data. Although inference is approximately 30% slower than the baseline, the performance gains present a compelling trade-off between accuracy and efficiency. The Wilcoxon signed-rank test indicates a significant difference between Mamba-HoME and other state-of-the-art methods, with a significance threshold of $p < 0.05$.

Table 4: Segmentation performance on the MVSeg test set. (*) indicates a pre-trained model.

Method	mDSC (%) ↑	mHD95 (mm) ↓
VoCo-B (*) [53]	84.3	17.2
Hermes [18]	83.5	13.3
Swin SMT [38]	83.4	17.2
VSmTrans [30]	84.4	17.4
uC 3DU-Net [22]	83.9	13.4
Swin UNETR [47]	84.4	13.1
SegMamba [54]	83.8	15.6
SuPreM (*) [28]	84.3	12.9
Mamba-HoME	84.8	12.6
Mamba-HoME (*)	**85.0**	**12.2**

Qualitative results. Figure 2 presents a qualitative comparison of our proposed Mamba-HoME method against the five top-performing baselines across three primary 3D medical imaging modalities: CT, MRI, and US. These modalities exhibit different organ contrasts, noise levels, and resolutions. Mamba-HoME demonstrates consistent improvements in segmentation quality across these scenarios. In the first row, it effectively handles small and closely located structures, showing precise boundary delineation while reducing common artifacts seen in baseline predictions. The second row highlights its capability to accurately segment organs of various shapes and sizes, even under low image quality conditions, with reduced susceptibility to over- or under-segmentation. The third row illustrates Mamba-HoME's robustness in handling noisy and low-resolution data, maintaining clear and anatomically accurate boundaries.

3.4 Ablation studies

In this section, we investigate the impact of several factors on the performance of Mamba-HoME: (1) the parameters of the HoME layer, including the number of experts in the first (E_1) and second (E_2) levels, the group size (K), and the number of slots per expert (S); (2) the effect of Dynamic Tanh normalization compared to Layer Normalization, specifically its influence on training and validation speed in SSMs and overall performance; and (3) the impact of the pre-trained model in a supervised learning approach. For each configuration, we evaluate the number of model parameters, GPU memory usage, and average DSC across three datasets[2]. More details on ablation studies can be found in the Appendix D.

[2]In these experiments, we use PANORAMA (PANO), AMOS-CT (AMOS), and FeTA.

Effect of the number of experts. We evaluate the impact of varying the number of experts at each encoder stage (i), where $i \in \{1, 2, 3, 4\}$, in a two-level HoME layer. For a fair comparison, we keep the group size constant ($K \in \{2048, 1024, 512, 256\}$) and set the number of slots to $S = 4$. Table 5 shows that the configuration with $E_1 \in \{4, 8, 12, 16\}$ experts at the first level and $E_2 \in \{8, 16, 24, 32\}$ experts at the second level achieves the best trade-off between segmentation performance and parameter efficiency. This setup also requires the fewest parameters and the lowest GPU memory usage.

Table 5: Quantitative segmentation performance of Mamba-HoME with varying numbers of experts at each encoder stage i.

#	Number of Experts (E)	Params (M) ↓	GPU (G) ↓	mDSC (%) ↑ PANO	AMOS	FeTA
1	$E_1 = [4, 8, 12, 16]$ $E_2 = [8, 16, 24, 32]$	**170.1**	**11.1**	**77.5**	**86.3**	**87.5**
2	$E_1 = [8, 16, 24, 48]$ $E_2 = [8, 16, 24, 48]$	277.8	12.2	76.3	86.2	87.3
3	$E_1 = [8, 16, 24, 48]$ $E_2 = [16, 32, 48, 96]$	359.2	12.9	76.0	85.9	87.4
4	$E_1 = [16, 32, 48, 64]$ $E_2 = [16, 32, 48, 64]$	367.0	13.0	**77.5**	86.1	87.4

Effect of the group size. We evaluate the impact of the group size at each encoder stage (i), where $i \in \{1, 2, 3, 4\}$, in a two-level HoME layer. For a fair comparison, we keep the number of experts constant ($E_1 \in \{4, 8, 12, 16\}$ and $E_2 \in \{8, 16, 24, 32\}$) and set the number of slots to $S = 4$. Table 6 shows that Mamba-HoME achieves optimal performance with group sizes $K \in \{2048, 1024, 512, 256\}$, while also minimizing GPU memory usage.

Table 6: Quantitative segmentation performance of Mamba-HoME with varying group sizes at each encoder stage i.

#	Group size (K)	Params (M) ↓	GPU (G) ↓	mDSC (%) ↑ PANO	AMOS	FeTA
1	[1024, 512, 256, 128]	**170.1**	**11.1**	77.2	86.1	87.4
2	[2048, 1024, 512, 256]	**170.1**	**11.1**	**77.5**	**86.3**	**87.5**
3	[2048, 1024, 512, 256]†	277.8	12.2	76.8	86.2	87.3
4	[4096, 2048, 1024, 512]	**170.1**	**11.1**	77.4	85.6	87.4

† The number of experts in the first and second levels of the HoME layer are equal ($E_1 = E_2 = [8, 16, 24, 48]$).

Effect of the number of slots per expert. We examine the impact of the number of slots (S) per expert (E) at each encoder stage (i), where $i \in \{1, 2, 3, 4\}$, in a two-level HoME layer. For a fair comparison, we keep the number of experts constant ($E_1 \in \{4, 8, 12, 16\}$ and $E_2 \in \{8, 16, 24, 32\}$) and the number of groups fixed ($K \in \{2048, 1024, 512, 256\}$).

Table 7 shows that Mamba-HoME achieves optimal performance at $S = 4$ across all evaluated datasets, representing a sweet spot for the number of slots per expert. Variations in the number of slots ($S \in \{1, 2, 8\}$) do not yield significant performance improvements, with other slot counts resulting in suboptimal performance while keeping the number of parameters and GPU memory constant.

Table 7: Quantitative segmentation performance of Mamba-HoME with varying numbers of slots per expert.

#	Slots (S)	mDSC (%) ↑ PANO	AMOS	FeTA
1	1	76.2	85.9	87.3
2	2	77.2	86.0	87.4
3	4	**77.5**	**86.3**	**87.5**
4	8	76.5	86.1	87.4

Effect of Dynamic Tanh normalization in SSMs. While DyT accelerates CNN- and Transformer-based architectures [58], we investigate its effectiveness in SSM-based architectures compared to Layer Normalization [7]. As shown in Table 8, segmentation performance remains largely unchanged, but DyT improves both training and inference speed by approximately 6% based on experimental runtime measurements[3].

Table 8: Quantitative segmentation performance of Mamba-HoME trained with Layer Normalization (LN) and Dynamic Tanh (DyT).

#	Dataset	LN	DyT	mDSC (%) ↑
1	PANO	✓	✗	77.4
2		✗	✓	**77.5**
3	FeTA	✓	✗	87.6
4		✗	✓	**87.7**
5	AMOS	✓	✗	**87.4**
6		✗	✓	87.3

Impact of supervised pre-training. We evaluate the efficacy of the proposed Mamba-HoME model, pre-trained under a supervised learning paradigm on publicly available datasets, including 8,788 CT and 616 MRI scans with voxel-wise annotations. The evaluation spans three primary 3D medical imaging modalities and various anatomical regions. Overall, Mamba-HoME outperforms state-of-the-art methods, surpassing existing approaches in both DSC and HD95 across all evaluated datasets. These quantitative results highlight the effectiveness of supervised pre-training in enhancing segmentation accuracy and robustness. A key feature of Mamba-HoME is its cross-modal generalization,

[3]For each dataset in this experiment, we adopt the same settings described in the Appendix B.

enabled by modality-agnostic feature representations. Pre-trained on CT and MRI scans, the model demonstrates superior adaptability to specialized tasks, such as fetal brain MRI segmentation in the FeTA dataset and 3D ultrasound mitral valve leaflet segmentation in the MVSeg dataset.

This adaptability highlights Mamba-HoME's ability to mitigate challenges posed by variations in modality, resolution, and clinical context. Moreover, its consistently high performance across heterogeneous datasets underscores its potential for practical deployment, where robust, modality-agnostic feature representations and precise segmentation are essential for scalable, real-world medical imaging applications. As shown in Figure 3, supervised pre-training significantly improves Mamba-HoME's performance compared to training from scratch or baseline methods, reducing artifacts and enhancing boundary segmentation for objects of varying sizes across the three primary 3D medical imaging modalities.

3.5 Generalizability analysis

To evaluate generalizability, we compare the proposed Mamba-HoME with several state-of-the-art networks. Specifically, we investigate four configurations on the AMOS dataset: (1) training solely on CT, (2) pre-training all models on CT and fine-tuning on MRI, (3) training solely on MRI, and (4) joint training on both CT and MRI. Table 2 shows that Mamba-HoME demonstrates superior generalizability across modalities compared to other models. Trained from scratch and further pre-trained on large-scale CT and MRI datasets, Mamba-HoME exhibits strong cross-modal generalizability to 3D ultrasound data, a modality with distinct challenges such as high noise and lower resolution. Leveraging robust, modality-agnostic feature representations, the pre-trained model adapts to 3D ultrasound via efficient fine-tuning, outperforming state-of-the-art methods in both DSC and boundary HD95 metrics, as shown in Table 4. Qualitative results in Figure 2 further illustrate its ability to handle ultrasound-specific artifacts. This cross-modal transferability highlights the model's versatility across diverse imaging modalities. Moreover, Mamba-HoME demonstrates strong generalizability to external datasets within the same modality, especially MSD Pancreas and in-house CT dataset for PDAC and pancreas segmentation, outperforming several state-of-the-art methods in both DSC and HD95 metrics (see Table 11). Detailed results for the generalizability analysis can be found in the Appendix F.

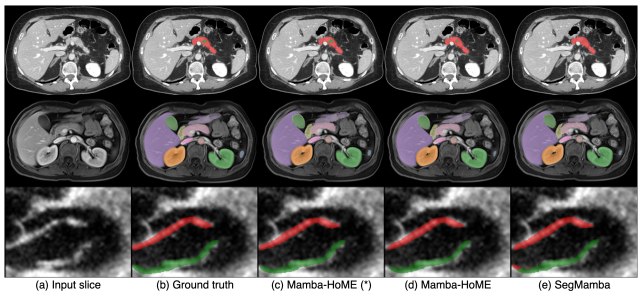

Figure 3: Qualitative segmentation results from top to bottom: CT, MRI, and 3D US. From left to right, each column shows the input slice, ground truth, our proposed pre-trained Mamba-HoME, Mamba-HoME trained from scratch, and the baseline SegMamba.

4 Conclusions

In this work, we introduce the Hierarchical Soft Mixture-of-Experts (HoME), a two-level token-routing MoE layer designed to efficiently capture local-to-global pattern hierarchies. We integrate HoME with Mamba in the Mamba-HoME architecture, enabling efficient long-sequence processing. Comprehensive experiments show that Mamba-HoME outperforms several state-of-the-art methods and generalizes well across the three primary 3D medical imaging modalities.

Limitations. Scalability to large-scale medical datasets (e.g., $>$10,000 scans) remains unexplored, limiting our understanding of Mamba-HoME's generalization across diverse image distributions. Although the model is pre-trained on a large multimodal dataset using supervised learning, its behavior under large-scale self-supervised learning (e.g., $>$200,000 scans) has not yet been studied. We identify this as a promising direction for future work to enhance Mamba-HoME's ability to capture complex patterns in unlabeled medical images. Key challenges include variations in image resolution, noise, contrast, field-of-view, acquisition techniques, and spatial-temporal dependencies.

Acknowledgments

We acknowledge the use of the HPC cluster at Helmholtz Munich for the computational resources used in this study. Ewa Szczurek acknowledges the support from the Polish National Science Centre SONATA BIS grant no. 2020/38/E/NZ2/00305.

References

[1] Tugba Akinci D'Antonoli, Lucas K Berger, Ashraya K Indrakanti, Nathan Vishwanathan, Jakob Weiss, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, et al. Totalsegmentator mri: Robust sequence-independent segmentation of multiple anatomic structures in mri. *Radiology*, 314(2):e241613, 2025.

[2] Nils Alves, Merlijn Schuurmans, Dominik Rutkowski, Derya Yakar, Ingfrid Haldorsen, Marjolein Liedenbaum, Anders Molven, Phillip Vendittelli, Geert Litjens, Jurgen Hermans, et al. The panorama study protocol: Pancreatic cancer diagnosis-radiologists meet ai. *Zenodo*, 2024.

[3] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022.

[4] Szymon Antoniak, Michał Krutul, Maciej Pióro, Jakub Krajewski, Jan Ludziejewski, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Marek Cygan, and Sebastian Jaszczur. Mixture of tokens: Continuous moe through cross-example aggregation. *Advances in Neural Information Processing Systems*, 37:103873–103896, 2024.

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[6] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[7] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[8] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

[9] Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *Advances in Neural Information Processing Systems*, 37:107547–107603, 2024.

[10] Patrick Carnahan, John Moore, Daniel Bainbridge, Mehdi Eskandari, Elvis CS Chen, and Terry M Peters. Deepmitral: Fully automatic 3d echocardiography segmentation for patient specific mitral valve modelling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 459–468. Springer, 2021.

[11] Ao Chang, Jiajun Zeng, Ruobing Huang, and Dong Ni. Em-net: Efficient channel and frequency learning with mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 266–275. Springer, 2024.

[12] Jieneng Chen, Yingda Xia, Jiawen Yao, Ke Yan, Jianpeng Zhang, Le Lu, Fakai Wang, Bo Zhou, Mingyan Qiu, Qihang Yu, et al. Cancerunit: Towards a single unified model for effective detection, segmentation, and diagnosis of eight major cancers using a large collection of ct scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21327–21338, 2023.

[13] Qian Chen, Lei Zhu, Hangzhou He, Xinliang Zhang, Shuang Zeng, Qiushi Ren, and Yanye Lu. Low-rank mixture-of-experts for continual medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 382–392. Springer, 2024.

[14] Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17346–17357, 2023.

[15] Tri Dao and Albert Gu. Transformers are ssms: generalized models and efficient algorithms through structured state space duality. In *Proceedings of the International Conference on Machine Learning*, pages 10041–10071, 2024.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *The Ninth International Conference on Learning Representations*, 2021.

[17] Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, Zhangyang Wang, et al. M^3vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. *Advances in Neural Information Processing Systems*, 35:28441–28457, 2022.

[18] Yunhe Gao. Training like a medical resident: Context-prior learning toward universal medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11194–11204, 2024.

[19] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[20] Joy Hsu, Jeffrey Gu, Gong Wu, Wah Chiu, and Serena Yeung. Capturing implicit hierarchical structure in 3d biomedical images with self-supervised hyperbolic representations. *Advances in Neural Information Processing Systems*, 34:5112–5123, 2021.

[21] Kai Hu, Jinhao Li, Yuan Zhang, Xiongjun Ye, and Xieping Gao. One-to-multiple: A progressive style transfer unsupervised domain-adaptive framework for kidney tumor segmentation. *Advances in Neural Information Processing Systems*, 37:24496–24522, 2024.

[22] Xingru Huang, Jian Huang, Tianyun Zhang, HE HONG, Shaowei Jiang, Yaoqi Sun, et al. Upping the game: How 2d u-net skip connections flip 3d segmentation. *Advances in Neural Information Processing Systems*, 37:87282–87309, 2025.

[23] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35:36722–36732, 2022.

[24] Mingjie Jiang and Bernard Chiu. A dual-stream centerline-guided network for segmentation of the common and internal carotid arteries from 3d ultrasound images. *IEEE Transactions on Medical Imaging*, 42(9):2690–2705, 2023.

[25] Yufeng Jiang and Yiqing Shen. M4oe: A foundation model for medical multimodal image segmentation with mixture of experts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 621–631. Springer, 2024.

[26] Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:131224–131246, 2024.

[27] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, 97:103285, 2024.

[28] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation? In *The Twelfth International Conference on Learning Representations*, 2024.

[29] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Lequan Yu, Yong Liang, Yizhou Yu, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-umamba†: Adapting mamba-based vision foundation models for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024.

[30] Tiange Liu, Qingze Bai, Drew A Torigian, Yubing Tong, and Jayaram K Udupa. Vsmtrans: A hybrid paradigm integrating self-attention and convolution for 3d medical image segmentation. *Medical Image Analysis*, 98:103295, 2024.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[32] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *The Fifth International Conference on Learning Representations*, 2017.

[33] Jiacheng Lu, Hui Ding, Shiyu Zhang, and Guoping Huo. M-net: Mri brain tumor sequential segmentation network via mesh-cast. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20116–20125, 2025.

[34] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[36] Kelly Payette, Hongwei Bran Li, Priscille de Dumast, Roxane Licandro, Hui Ji, Md Mahfuzur Rahman Siddiquee, Daguang Xu, Andriy Myronenko, Hao Liu, Yuchen Pei, et al. Fetal brain tissue annotation and segmentation challenge results. *Medical Image Analysis*, 88:102833, 2023.

[37] Kelly Payette, Céline Steger, Roxane Licandro, Priscille De Dumast, Hongwei Bran Li, Matthew Barkovich, Liu Li, Maik Dannecker, Chen Chen, Cheng Ouyang, et al. Multi-center fetal brain tissue annotation (feta) challenge 2022 results. *IEEE Transactions on Medical Imaging*, 2024.

[38] Szymon Płotka, Maciej Chrabaszcz, and Przemyslaw Biecek. Swin smt: Global sequential modeling for enhancing 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 689–698. Springer, 2024.

[39] Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *The Twelfth International Conference on Learning Representations*, 2024.

[40] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Yucheng Tang, Alan L Yuille, Zongwei Zhou, et al. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems*, 36:36620–36636, 2023.

[41] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

[42] 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.

[43] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 556–564. Springer, 2015.

[44] Tomasz Szczepański, Michal K Grzeszczyk, Szymon Płotka, Arleta Adamowicz, Piotr Fudalej, Przemysław Korzeniowski, Tomasz Trzciński, and Arkadiusz Sitek. Let me decode you: Decoder conditioning with tabular data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 228–238. Springer, 2024.

[45] Tomasz Szczepański, Szymon Płotka, Michal K Grzeszczyk, Arleta Adamowicz, Piotr Fudalej, Przemysław Korzeniowski, Tomasz Trzciński, and Arkadiusz Sitek. Gepar3d: Geometry prior-assisted learning for 3d tooth segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 218–228. Springer, 2025.

[46] Fenghe Tang, Bingkun Nian, Yingtai Li, Zihang Jiang, Jie Yang, Wei Liu, and S Kevin Zhou. Mambamim: Pre-training mamba with state space token interpolation and its application to medical image segmentation. *Medical Image Analysis*, 103:103606, 2025.

[47] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.

[48] Arvind M Vepa, Zukang Yang, Andrew Choi, Jungseock Joo, Fabien Scalzo, and Yizhou Sun. Integrating deep metric learning with coreset for active learning in 3d segmentation. *Advances in Neural Information Processing Systems*, 37:71643–71671, 2024.

[49] Guoan Wang, Jin Ye, Junlong Cheng, Tianbin Li, Zhaolin Chen, Jianfei Cai, Junjun He, and Bohan Zhuang. Sam-med3d-moe: Towards a non-forgetting segment anything model via mixture of experts for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer, 2024.

[50] Hualiang Wang, Yiqun Lin, Xinpeng Ding, and Xiaomeng Li. Tri-plane mamba: Efficiently adapting segment anything model for 3d medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 636–646. Springer, 2024.

[51] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

[52] Boqian Wu, Qiao Xiao, Shiwei Liu, Lu Yin, Mykola Pechenizkiy, Decebal Constantin Mocanu, Maurice Keulen, and Elena Mocanu. E2enet: Dynamic sparse feature fusion for accurate and efficient 3d medical image segmentation. *Advances in Neural Information Processing Systems*, 37:118483–118512, 2025.

[53] Linshan Wu, Jiaxin Zhuang, and Hao Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22873–22882, 2024.

[54] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–588. Springer, 2024.

[55] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36:41693–41706, 2023.

[56] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

[57] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2019.

[58] Jiachen Zhu, Xinlei Chen, Kaiming He, Yann LeCun, and Zhuang Liu. Transformers without normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[59] Xingkui Zhu, Yiran Guan, Dingkang Liang, Yuchao Chen, Yuliang Liu, and Xiang Bai. Moe jetpack: From dense checkpoints to adaptive mixture of experts for vision tasks. *Advances in Neural Information Processing Systems*, 37:12094–12118, 2024.

[60] Veronika A Zimmer, Alberto Gomez, Emily Skelton, Robert Wright, Gavin Wheeler, Shujie Deng, Nooshin Ghavami, Karen Lloyd, Jacqueline Matthew, Bernhard Kainz, et al. Placenta segmentation in ultrasound imaging: Addressing sources of uncertainty and limited field-of-view. *Medical Image Analysis*, 83:102639, 2023.

Appendix

Table of Contents

A Datasets

A core objective of this work is to assess the robustness and generalizability of Mamba-HoME. To this end, we conduct a comprehensive evaluation using datasets from three primary 3D medical imaging modalities: CT, MRI, and US. This approach provides a full spectrum from low to high-resolution images with varying levels of noise and artefacts. Table 9 shows an overview of the datasets used for pre-training, training, fine-tuning, and testing.

Table 9: An overview of the datasets used for pre-training, training, fine-tuning, and testing. These datasets, spanning three modalities (CT, MRI, and US), cover diverse anatomical structures and lesions. Please note that all datasets configured in training mode were utilized for fine-tuning.

No.	Dataset	Modality	Body part	Mode	Label type	Pre-training	Train	Test
1.	AbdomenAtlas 1.1 [40, 27, 28]	CT	Abdomen	Pre-training	Voxel-wise	8,788 (-474)[4]	-	-
2.	TotalSegmentator MRI [1]	MRI	Whole-body	Pre-training	Voxel-wise	616	-	-
3.	PANORAMA [2]	CT	Abdomen	Training	Voxel-wise	-	1,964	334
4.	AMOS [23]	CT-MRI	Abdomen	Training	Voxel-wise	-	240	120
5.	FeTA 2022 [37]	MRI	Fetal brain	Training	Voxel-wise	-	120	-
6.	MVSeg [10]	US	Heart	Training	Voxel-wise	-	110	-
7.	In-house CT	CT	Abdomen	Test	Voxel-wise	-	-	60
					Total	**9,404**	**2,434**	**514**

A.1 Pre-training

For pre-training, we utilize two manually annotated, publicly available large-scale datasets with two modalities: AbdomenAtlas 1.1 [27, 40, 28], which includes CT scans, and TotalSegmentator MRI [1], which includes MRI scans.

AbdomenAtlas. AbdomenAtlas 1.1 dataset consists of 9,262 CT scans with manually voxel-wise annotated 25 anatomical structures. The dataset consists of a mix of non-contrast, arterial, portal-venous, and delayed phases. Each CT image consists of 24∼2,572 slices, with a resolution ranging from 188×79 to 971×651 pixels. The voxel spatial resolution ranges ([0.38∼1.5]×[0.38∼3.0]×[0.3∼8.0]) mm^3 with a mean of 0.84 × 0.84 × 2.4 mm^3. From AbdomenAtlas 1.1, we excluded overlap cases from the PANORAMA dataset, including 194 MSD Pancreas cases, 43 NIH Pancreas cases, and 200 training cases from the AMOS-CT dataset.

TotalSegmentator MRI. The TotalSegmentator MRI dataset consists of 616 MRI images with 50 anatomical structures manually annotated at the voxel level. For pre-training, we select 22 classes that match those in the AbdomenAtlas dataset. Each MRI scan consists of 5∼1,915 slices. The voxel spatial resolution ranges ([0.17∼20.0]×[0.17∼25.0]×[0.17∼28.0]) mm^3 with a mean of 1.28 × 1.86 × 2.81 mm^3. Table 10 shows the class map from both modalities used for pre-training the model.

A.2 Training, fine-tuning, and test

To validate the efficiency of the Mamba-HoME, we conduct comprehensive experiments on both publicly available and in-house datasets across three primary modalities. For CT, we use PANORAMA [2], AMOS-CT [23], and a private dataset for segmentation and diagnosis of pancreatic ductal adenocarcinoma (PDAC) in abdominal CT. For MRI, we use AMOS-MRI [23], and FeTA 2022, which includes brain fetal MRI [36, 37]. For US, we use MVSeg [10].

PANORAMA. The PANORAMA dataset [2] comprises 2,238 multi-center contrast-enhanced computed tomography (CECT) scans acquired in a single portal-venous phase. It includes 1,964 newly acquired scans from five European centers (the Netherlands, Norway, and Sweden), as well as publicly available data from two medical centers in the United States, namely NIH [43] and MSD Pancreas [3]. The dataset consists of 676 PDAC (pancreatic ductal adenocarcinoma) and 1,562 non-PDAC cases, respectively. While the dataset contains six voxel-wise labels, PDAC lesion, veins, arteries, pancreatic parenchyma, pancreatic duct, and common bile duct, we use only the PDAC lesion and merge the pancreatic parenchyma and pancreatic duct classes into a single *pancreas* label. Each CT

[4]Please note that, for a fair comparison, we excluded 200 cases from AMOS, 194 from MSD Pancreas, and 80 from NIH, respectively. These scans correspond to the original AMOS dataset and partly to the PANORAMA dataset.

Table 10: Class mapping for anatomical structures in CT scans (AbdomenAtlas) and MRI scans (TotalSegmentator MRI).

Index	Class	Index	Class
0	Background	13	Celiac trunk
1	Aorta	14	Colon
2	Gall bladder	15	Duodenum
3	Kidney (left)	16	Esophagus
4	Kidney (right)	17	Femur (left)
5	Liver	18	Femur (right)
6	Pancreas	19	Hepatic vessel
7	Postcava	20	Intestine
8	Spleen	21	Lung (left)
9	Stomach	22	Lung (right)
10	Adrenal gland (left)	23	Portal vein & splenic vein
11	Adrenal gland (right)	24	Prostate
12	Bladder	25	Rectum

image consists of 37∼1,572 slices, with a resolution ranging from 512×512 to 1024×1024 pixels. The voxel spatial resolution ranges ([0.31∼1.03]×[0.31∼1.03]×[0.45∼5.0]) mm^3 with a mean of $0.75 \times 0.75 \times 2.0$ mm^3.

AMOS. The dataset consists of a total of 600 scans (500 CT and 100 MRI). However, the ground truth is available only for the training and validation sets, which include 240 and 120 scans, respectively. Each scan contains 15 anatomical structures, manually annotated at the voxel level. For the training and evaluation of AMOS-CT, we use 200 and 100 scans for training and testing, respectively. For AMOS-MRI, we use 40 and 20 scans for training and testing, respectively. Each CT scan consists of 68∼353 slices, with a voxel spatial resolution ranging from ([0.45∼1.07]×[0.45∼1.07]×[1.25∼5.0]) mm^3, with a mean of 0.70×0.70×4.20 mm^3. Each MRI scan consists of 60∼168 slices, with a voxel spatial resolution ranging from ([0.70∼1.95]×[0.70∼1.95]×[1.09∼3.0]) mm^3, with a mean of 1.10×1.10×2.46 mm^3.

FeTA 2022. The dataset comprises 120 cases of MRI scans for fetal brain tissue segmentation, collected from two prominent medical centers: the University Children's Hospital Zurich and the Medical University of Vienna. The dataset includes 80 cases from Zurich and 40 from Vienna. It features seven manually annotated tissues, with voxel-wise annotations provided for each: external cerebrospinal fluid, gray matter, white matter, ventricles, cerebellum, deep gray matter, and the brainstem. Each fetal MRI scan consists of 256 slices, with a voxel spatial resolution ranging from ([0.43∼1.0]×[0.43∼1.0]×[0.43∼1.0]) mm^3, with a mean of 0.67×0.67×0.67 mm^3.

MVSeg. The dataset [10] consists of 175 transesophageal echocardiography (TEE) 3D US scans for mitral valve segmentation. Each scan consists of voxel-wise annotations provided for the posterior leaflet and the anterior leaflet. Data was acquired at King's College Hospital, London, UK. Each 3D US image consists of 208 slices. The voxel spatial resolution ranges ([0.20∼0.63]×[0.31∼0.90]×[0.13∼0.39]) mm^3 with a mean of $0.38 \times 0.56 \times 0.23$ mm^3.

In-house CT. 60 CECT scans of histopathology-confirmed PDAC cases, acquired from 40 medical centers, each containing 68∼875 slices with a voxel spatial resolution ranging ([0.47∼0.98]×[0.47∼0.98]×[0.5∼5.0]) mm^3. These scans contain voxel-wise annotations of both the pancreas and the PDAC. The diameter size of the PDAC ranges 1.25∼6.06 cm, with a mean of 3.45±1.20 cm. All test CECT scans have a resolution of 512 × 512 pixels. Two junior radiologists annotated the in-house data using 3D Slicer, which was reviewed by two domain experts with more than 30 years of experience.

B Experimental setup

B.1 Pre-training

AbdomenAtlas and TotalSegmentator MRI. Each CT and MRI scan is resampled into a spacing of $0.8 \times 0.8 \times 3.0$ mm^3 resolution. Before training, we randomly split the data in an 85:15 ratio, ensuring a consistent number of CT and MRI scans in both the training and validation sets. All CT scans are clipped to the window of the Hounsfield Unit of $[-175, 250]$, while MRI scans are clipped to $[0, 1000]$. Both modalities are linearly scaled to $[0, 1]$. We train Mamba-HoME with a patch size of $96 \times 96 \times 96$ for 800 epochs. During the inference stage, we use a sliding-window algorithm with a patch size of $96 \times 96 \times 96$ and an overall of 0.5 with a Gaussian filter. The pre-training took around 7 days.

B.2 Training and fine-tuning

PANORAMA. Each CT scan is resampled into the spacing of $0.8 \times 0.8 \times 3$ mm^3 resolution. All CT scans are clipped to the window of the Hounsfield Unit of $[-175, 250]$ and linearly scaled to $[0, 1]$. During training, we randomly select 1,571 scans (80%) for training and 393 scans (20%) for validation sets, maintaining the distribution ratio of PDAC to non-PDAC cases. During the inference stage, we use a sliding-window algorithm with a patch size of $192 \times 192 \times 48$ and an overlap of 0.5 with a Gaussian filter. We train the model for 500 epochs.

AMOS-CT and AMOS-MRI. We resample each **CT** and **MRI** scan to the $1.5 \times 1.5 \times 1.5$ mm^3 isotropic resolution. All CT scans are clipped to the window of the Hounsfield Unit of $[-175, 250]$, while MRI scans are clipped to $[0, 1000]$. Both modalities are linearly scaled to $[0, 1]$. We randomly split an original training dataset into train and validation subsets with 80:20 ratio. During the training, we crop random patches of $128 \times 128 \times 128$, and a sliding window algorithm with a default overlap of 0.5 is performed for validation. During training, we employ on-the-fly augmentations, including scaling, rotation, flipping, adjusting brightness and contrast, and Gaussian smoothness and noise.

FeTA 2022. We resample each MRI volume into $0.8 \times 0.8 \times 0.8$ mm^3 isotropic resolution. We clip each MRI scan value of $[0, 1000]$ and linearly scale to $[0, 1]$. We train the models using a five-fold cross-validation strategy for 300 epochs each. During training, we crop random patches of $128 \times 128 \times 128$, and we employ on-the-fly augmentations, including scaling, rotation, flipping, adjusting brightness and contrast, Gaussian smoothness and noise, and affine transformations.

MVSeg. We resample each US scan into $0.5 \times 0.5 \times 0.5$ mm^3 isotropic resolution. We clip each US scan value of $[0, 255]$ and linearly scale to $[0, 1]$. We train the models using an original train, valid, and test split, including 105, 30, and 40 scans, respectively. We use 500 epochs for each training. During training, we crop random patches of $128 \times 128 \times 128$, and we employ on-the-fly augmentations, including scaling, rotation, flipping, adjusting brightness and contrast, Gaussian smoothness and noise, and affine transformations.

B.3 Evaluation metrics

Dice Similarity Coefficient (DSC). The DSC is a commonly used metric to evaluate the segmentation performance of a model in multi-class settings, especially in medical imaging. It measures the overlap between predicted and ground truth segmentations, providing an aggregate assessment across multiple classes. Given a segmentation task with C classes, let $\mathbf{p}_i \in \mathbb{R}^C$ and $\mathbf{g}_i \in \mathbb{R}^C$ be the one-hot encoded predicted and ground truth vectors at voxel i, respectively. The DSC for class c is computed as:

$$DSC_c = \frac{2 \sum_i p_{i,c} g_{i,c}}{\sum_i p_{i,c} + \sum_i g_{i,c}}, \tag{16}$$

where $p_{i,c}$ and $g_{i,c}$ represent the predicted and ground truth binary masks for class c at voxel i, respectively. The mean DSC (mDSC) across all C classes is computed as follows:

$$\text{mDSC} = \frac{1}{C} \sum_{c=1}^{C} DSC_c. \tag{17}$$

This metric ensures that each class contributes equally to the final score, regardless of class imbalance.

95th percentile Hausdorff Distance (HD95). The HD95 is a commonly used metric to evaluate the spatial similarity between predicted and ground truth segmentations. It measures the worst-case boundary discrepancy between two sets but is made more robust by considering the 95th percentile instead of the maximum distance. Given a segmentation task with C classes, let S_c^P and S_c^G denote the sets of boundary points for the predicted and ground truth segmentations for class c, respectively. The directed Hausdorff distance from S_c^P to S_c^G is defined as:

$$d_{\mathrm{H}}(S_c^P, S_c^G) = \max_{p \in S_c^P} \min_{g \in S_c^G} \|p - g\|_2, \tag{18}$$

where $\|p - g\|_2$ denotes the Euclidean distance between points p and g. The bi-directional Hausdorff distance is given by:

$$d_{\mathrm{H}}(S_c^P, S_c^G) = \max\left(\max_{p \in S_c^P} \min_{g \in S_c^G} \|p - g\|_2, \max_{g \in S_c^G} \min_{p \in S_c^P} \|g - p\|_2 \right). \tag{19}$$

Instead of using the maximum, the 95th percentile Hausdorff distance is computed to mitigate the influence of outliers:

$$HD95_c = \text{percentile}_{95}\left(\{ d_{\mathrm{H}}(S_c^P, S_c^G), d_{\mathrm{H}}(S_c^G, S_c^P) \} \right). \tag{20}$$

The mean HD95 (mHD95) across all classes is then given by:

$$\text{mHD95} = \frac{1}{C} \sum_{c=1}^{C} HD95_c. \tag{21}$$

This metric is widely used in medical image segmentation to quantify boundary errors while reducing the sensitivity to small outlier deviations.

Sensitivity and Specificity. The sensitivity and specificity are crucial metrics for evaluating the performance of a model in detecting the presence of a condition at the patient level. In this setting, a segmentation model processes medical scans and outputs a binary classification for each patient: either *positive* (presence of the condition) or *negative* (absence of the condition).

Given a dataset of patients, let TP, FP, FN, and TN denote the number of true positives, false positives, false negatives, and true negatives, respectively. The sensitivity (also known as recall or true positive rate) is defined as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{22}$$

where sensitivity measures the proportion of correctly identified positive patients out of all actual positive patients. The specificity (true negative rate) is given by:

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{23}$$

where specificity quantifies the proportion of correctly identified negative patients out of all actual negative patients.

These metrics provide a comprehensive assessment of the model's ability to detect the condition while avoiding false alarms, which is critical for clinical decision-making.

Number of parameters. The total number of trainable parameters in a neural network can be computed by summing the parameters across all layers and levels, as follows:

$$\text{Parameters} = \|\Theta\|_0 + \sum_{l=1}^{L} \sum_{k=1}^{L-l} \sum_{i=1}^{C_l} \sum_{j=1}^{C_k} A_{l,k}^{i,j} \|\theta_{l,k}^{i,j}\|_0, \tag{24}$$

where Θ represents the set of parameters from the backbone, L is the total number of levels or layers in the network, C_l and C_k represent the number of input and output channels at each level l and k, respectively, $A_{l,k}^{i,j}$ is a binary matrix indicating the presence of a weight connection between input channel i at level l and output channel j at level k, $\|\theta_{l,k}^{i,j}\|_0$ represents the count of non-zero weights in the connection between input channel i and output channel j.

This formulation takes into account the layer-wise parameters while incorporating the structured sparsity of weights within the network.

C Additional experimental results

Table 11: Quantitative segmentation results on PANORAMA test sets, including NIH, MSD Pancreas, and one in-house set. We report mDSC (%) and mHD95 (mm) for the pancreas and PDAC. Please note, NIH dataset consists only of healthy controls. Patient-level PDAC detection is evaluated using sensitivity (%) and specificity (%). The best results are **bolded**, while the second-best are underlined. (*) indicates pre-trained model. Please note, SuPreM was partially pre-trained on a subset of the test set, including the MSD Pancreas and NIH datasets, which may result in an unfair comparison.

Method	mDSC (%) ↑							mHD95 (mm) ↓		PDAC Detection ↑	
	NIH	MSD Pancreas		In-house		Overall		Overall		Overall (%)	
	Pancreas	PDAC	Pancreas	PDAC	Pancreas	PDAC	Pancreas	Pancreas	PDAC	Sensitivity	Specificity
VoCo-B (*)	90.1±5.0	38.7±28.9	86.5±7.2	43.3±24.1	80.3±11.7	40.5±27.2	86.3±8.4	3.0±4.8	271.5±144.2	86.1	91.6
Hermes	92.0±2.8	40.0±32.8	86.5±8.9	58.3±25.6	80.8±11.3	48.2±30.6	87.8±8.0	2.7±4.9	23.9±60.5	87.3	94.0
Swin SMT	91.9±2.8	44.0±32.6	87.2±6.0	58.6±24.2	79.8±11.0	49.4±30.6	87.0±7.8	4.8±23.9	92.1±124.7	88.6	92.8
VSmTrans	92.0±2.4	47.1±30.6	87.3±6.5	56.1±26.9	80.5±11.4	50.3±29.6	87.2±7.9	3.5±15.8	82.1±65.4	89.9	90.4
SegMamba	92.7±2.6	47.9±32.5	88.8±6.4	54.0±29.6	81.9±10.5	49.7±31.5	**88.5±7.6**	4.3±22.1	36.3±94.7	84.8	**95.2**
uC 3DU-Net	92.6±2.6	49.0±31.2	**88.9±6.7**	56.8±26.5	80.2±11.8	52.0±29.7	88.2±8.3	2.9±4.1	18.2±21.0	80.7	89.2
Swin UNETR	91.9±2.1	38.8±28.8	87.4±5.7	58.7±23.3	81.3±9.8	46.3±28.5	87.4±7.0	5.9±25.0	187.1±112.3	87.3	90.4
SuPreM (*)	92.3±2.7	46.0±33.3	88.6±5.1	61.4±24.0	82.0±10.2	51.7±31.2	88.3±6.8	2.4±4.3	9.0±13.0	87.3	89.2
Mamba-HoME	**92.9±2.2**	51.2±30.5	88.4±7.0	60.8±26.8	81.7±11.2	54.8±29.5	88.3±8.1	2.3±4.2	8.4±12.1	90.4	92.8
Mamba-HoME (*)	92.6±2.4	**53.6±31.9**	88.5±6.8	**61.7±25.7**	**82.3±10.3**	**56.7±29.9**	**88.5±7.6**	**1.9±3.1**	**6.2±10.2**	**92.8**	**95.2**

Table 12: Quantitative segmentation results of 5-fold cross-validation on the FeTA 2022 dataset. We report the mDSC (%) and HD95 (mm) for each class and across all folds. Results are provided for external cerebrospinal fluid (eCSF), gray matter (GM), white matter (WM), ventricles (V), cerebellum (C), deep gray matter (dGM), and brainstem (B). The best results are **bolded**, while the second-best are underlined. (*) indicates pre-trained model.

Method	DSC (%) ↑							mDSC (%) ↑	HD95 (mm) ↓
	eCSF	GM	WM	V	C	dGM	B		
VoCo-B (*)	78.9±6.2	73.4±3.1	90.2±1.8	87.5±1.2	87.0±2.9	87.6±1.6	83.8±2.6	86.0±1.8	4.0±3.2
Swin SMT	78.3±5.7	72.6±2.8	90.0±1.8	86.2±2.3	86.6±1.1	87.9±1.9	83.8±2.0	85.6±1.3	2.4±0.3
Hermes	79.3±6.0	74.2±2.8	90.7±1.4	87.6±1.3	87.1±2.3	88.8±1.8	84.7±2.4	86.5±1.5	4.0±3.4
SegMamba	79.6±5.7	73.7±3.1	90.6±1.5	87.1±1.7	87.5±1.9	88.8±1.6	85.2±1.9	86.5±1.2	5.8±1.9
uC 3DU-Net	79.1±6.1	73.2±2.9	90.3±1.5	87.4±1.2	86.0±2.8	88.0±1.8	83.4±2.6	85.9±1.7	3.5±2.1
SuPreM (*)	78.0±6.1	71.7±3.4	89.8±1.7	86.8±1.8	85.9±3.6	87.1±1.4	83.5±2.7	85.3±1.8	3.6±2.3
VSmTrans	79.0±5.7	73.1±3.1	90.6±1.5	87.2±1.7	87.2±0.9	88.3±1.7	84.2±2.1	86.1±1.2	2.3±0.3
Mamba-HoME	80.4±5.8	75.3±3.1	91.2±1.5	88.2±1.3	88.8±1.7	**89.4±1.8**	**86.6±1.5**	87.5±1.2	2.1±0.2
Mamba-HoME (*)	**80.6±4.3**	**76.1±2.6**	**91.9±1.1**	**88.8±0.8**	**90.0±0.9**	88.8±1.9	86.1±1.2	**87.7±1.0**	**2.0±0.2**

Table 13: Quantitative segmentation results on MVSeg test set. We report mDSC (%) and mHD95 (mm) for the posterior leaflet and the anterior leaflet. The best results are **bolded**, while the second-best are underlined. (*) indicates pre-trained model.

Method	DSC (%) ↑		mDSC (%) ↑	mHD95 (mm) ↓
	Posterior leaflet	Anterior leaflet		
Swin SMT	82.2±8.2	84.5±4.7	83.4±6.8	17.2±21.9
Hermes	82.9±6.8	84.1±5.5	83.5±6.2	13.3±13.8
uC 3DU-Net	82.9±6.8	85.0±4.3	83.9±5.8	13.4±15.7
SuPreM (*)	83.2±7.6	85.4±4.7	84.3±6.4	12.9±14.6
VoCo-B (*)	83.3±7.3	85.3±3.9	84.3±5.9	17.2±22.0
Swin UNETR	83.8±5.9	85.1±4.6	84.4±5.3	13.1±12.3
SegMamba	83.0±5.8	84.7±4.1	83.8±5.1	15.6±16.4
VSmTrans	83.4±6.5	85.4±3.7	84.4±5.5	17.4±24.4
Mamba-HoME	84.0±5.8	85.7±3.8	84.8±5.1	12.6±12.6
Mamba-HoME (*)	**84.2±5.8**	**86.0±3.7**	**85.0±4.9**	**12.2±12.3**

D Ablation studies

We present detailed quantitative results from our ablation studies, analyzing three key factors: (1) the number of experts per stage; (2) the group size; and (3) the number of slots assigned to each expert.

D.1 Effect of the number of experts

For a fair comparison, we keep the group size constant ($K \in \{2048, 1024, 512, 256\}$) and set the number of slots to $S = 4$.

Table 14: Quantitative segmentation results of the effect of the number of experts on the PANORAMA test sets. We report mDSC (%) and mHD95 (mm) for PDAC and Pancreas classes across datasets, including NIH, MSD, in-house and overall. Results are provided for pancreas in NIH, PDAC and pancreas in MSD, in-house, and overall. Note: NIH dataset contains only the Pancreas class. The best results are **bolded**, while the second-best are underlined. (*) indicates a doubled number of experts in the second level.

Experts (E)	NIH	MSD		In-house		Overall		mDSC	mHD95
	Pancreas	PDAC	Pancreas	PDAC	Pancreas	PDAC	Pancreas	(%) ↑	(mm) ↓
[4, 8, 12, 16]*	92.6±2.4	**51.2±30.4**	88.4±7.0	**60.8±26.8**	81.7±11.2	**54.8±29.5**	88.3±8.1	**77.5±21.2**	**6.9±12.1**
[8, 16, 24, 48]	92.6±2.5	49.4±32.1	88.2±7.2	59.6±28.3	80.6±13.2	53.3±31.2	87.9±8.8	76.3±25.0	8.4±14.2
[8, 16, 24, 48]*	92.1±3.5	47.9±34.3	88.0±9.1	57.4±31.2	80.5±15.4	52.2±32.2	87.5±8.9	75.8±26.2	9.5±15.2
[16, 32, 48, 64]	**92.7±2.2**	51.0±31.2	**88.5±7.1**	60.4±28.1	**81.9±11.0**	54.5±28.5	**88.4±8.2**	**77.5±21.5**	7.2±14.1

Table 15: Quantitative segmentation results of AMOS-CT validation dataset. We report mDSC (%) and mHD95 (mm) for each organ. Organs include Spleen (Sp), Right Kidney (RK), Left Kidney (LK), Gallbladder (GB), Esophagus (Es), Liver (Li), Stomach (St), Aorta (Ao), Inferior Vena Cava (IVC), Pancreas (Pa), Right Adrenal Gland (RAG), Left Adrenal Gland (LAG), Duodenum (Du), Bladder (Bl), and Prostate/Uterus (Pr/Ut). The best results are **bolded**, while the second-best are underlined. (*) indicates a doubled number of experts in the second level.

Experts (E)	DSC (%) ↑								
	Sp	RK	LK	GB	Es	Li	St	Ao	IVC
[4, 8, 12, 16]*	95.7±4.1	95.3±2.8	**95.1±3.2**	81.4±20.3	81.5±11.9	93.1±12.1	88.7±13.6	**93.6±4.8**	**88.4±4.4**
[8, 16, 24, 48]	95.7±4.2	93.7±6.4	94.0±7.0	81.0±20.8	81.5±13.1	95.3±7.0	**89.3±13.6**	93.2±4.9	**88.4±4.4**
[8, 16, 24, 48]*	94.7±6.0	**95.4±2.7**	95.0±3.6	**81.6±20.9**	81.4±12.3	95.4±6.4	87.0±15.0	93.4±4.8	87.8±5.1
[16, 32, 48, 64]	**95.8±3.7**	93.8±4.3	94.8±6.0	80.0±22.4	83.0±10.1	**95.7±4.8**	87.5±15.0	**93.6±4.6**	**88.4±4.4**

Experts (E)							mDSC	mHD95
	Pa	RAG	LAG	Du	Bl	Pr/Ut	(%) ↑	(mm) ↓
[4, 8, 12, 16]*	83.6±11.1	74.1±11.7	**75.2±11.1**	77.4±14.1	**86.3±17.7**	83.2±13.8	**86.3±13.5**	**45.0±85.1**
[8, 16, 24, 48]	83.4±11.2	73.8±12.0	75.0±13.0	**77.6±14.3**	87.6±16.8	**83.3±15.5**	86.2±13.9	49.2±91.2
[8, 16, 24, 48]*	83.4±11.9	73.8±11.6	74.4±12.4	77.0±12.6	86.0±17.1	82.2±16.3	85.9±13.9	55.6±75.2
[16, 32, 48, 64]	**83.8±11.0**	**75.0±11.3**	74.5±12.5	77.5±13.5	85.7±17.3	82.4±16.9	86.1±13.9	50.3±84.2

Table 16: 5-fold cross-validation on the FeTA 2022 dataset of the varying number of experts of HoME layer at each encoder stage. We report mDSC (%) and mHD95 (mm) for each class and across all folds. Quantitative segmentation results are provided for external cerebrospinal fluid (eCSF), gray matter (GM), white matter (WM), ventricles (V), cerebellum (C), deep gray matter (dGM), and brainstem (B). The best results are **bolded**, while the second-best are underlined. (*) indicates a doubled number of experts in the second level.

Experts (E)	DSC (%) ↑							mDSC	HD95
	eCSF	GM	WM	V	C	dGM	B	(%) ↑	(mm) ↓
[4, 8, 12, 16]*	80.6±5.9	75.3±3.1	91.2±1.4	88.2±1.3	88.3±1.5	89.3±1.8	86.7±1.6	**87.5±1.2**	**2.2±0.4**
[8, 16, 24, 48]	80.4±6.2	75.1±3.4	91.1±1.5	88.3±1.2	**88.5±1.2**	88.6±2.1	86.6±1.5	87.2±1.4	2.4±0.5
[8, 16, 24, 48]*	**80.7±5.7**	**75.5±3.2**	91.1±1.5	88.4±1.2	**88.5±1.8**	**89.4±1.4**	**86.9±1.6**	87.4±1.5	2.3±0.4
[16, 32, 48, 64]	80.5±5.7	75.4±3.4	**91.5±1.2**	**88.5±1.6**	88.0±1.4	88.9±1.7	86.5±1.7	87.4±1.6	2.3±0.5

D.2 Effect of the group size

For a fair comparison, we keep the number of experts constant ($E_1 \in \{4, 8, 12, 16\}$ and $E_2 \in \{8, 16, 24, 32\}$) and set the number of slots to $S = 4$.

Table 17: Quantitative segmentation results of the effect of the number of slots per expert on the PANORAMA test sets. We report mDSC (%) and mHD95 (mm) for PDAC and Pancreas classes across datasets, including NIH, MSD, in-house and overall. Results are provided for pancreas in NIH, PDAC and pancreas in MSD, in-house, and overall. The best results are **bolded**, while the second-best are underlined. (*) indicates a doubled number of experts in the second level.

Group size (K)	NIH	MSD		In-house		Overall		mDSC (%) ↑	mHD95 (mm) ↓
	Pancreas	PDAC	Pancreas	PDAC	Pancreas	PDAC	Pancreas		
[1024, 512, 256, 128]	92.5±2.6	50.4±30.2	88.2±6.4	60.5±27.1	81.5±10.5	54.2±30.5	88.1±8.0	77.2±24.7	9.2±5.1
[2048, 1024, 512, 256]	**92.9±2.2**	51.2±30.5	**88.4±7.0**	**60.8±26.8**	**81.7±11.2**	54.8±29.5	**88.3±8.1**	**77.5±23.8**	**6.9±4.0**
[2048, 1024, 512, 256]*	92.7±2.7	50.3±31.5	88.3±7.0	56.3±29.1	**81.7±10.9**	52.6±30.8	88.2±8.0	76.8±25.0	11.2±5.7
[4096, 2048, 1024, 512]	92.5±2.5	**51.9±30.2**	88.1±6.6	60.6±27.0	81.5±10.7	**55.2±29.3**	88.0±7.7	77.4±23.5	7.2±4.2

Table 18: Quantitative segmentation results of AMOS-CT validation dataset. We report mDSC (%) and mHD95 (mm) for each organ. Organs include Spleen (Sp), Right Kidney (RK), Left Kidney (LK), Gallbladder (GB), Esophagus (Es), Liver (Li), Stomach (St), Aorta (Ao), Inferior Vena Cava (IVC), Pancreas (Pa), Right Adrenal Gland (RAG), Left Adrenal Gland (LAG), Duodenum (Du), Bladder (Bl), and Prostate/Uterus (Pr/Ut). The best results are **bolded**, while the second-best are underlined. (*) indicates a doubled number of experts in the second level.

Group size (K)	DSC (%) ↑								
	Sp	RK	LK	GB	Es	Li	St	Ao	IVC
[4096, 2048, 1024, 512]	95.7±3.8	95.1±3.3	95.5±2.7	81.2±20.3	82.0±10.8	89.0±18.6	86.5±16.5	93.5±4.1	87.8±5.7
[2048, 1024, 512, 256]	**95.8±3.7**	93.8±4.3	94.8±6.0	80.0±22.4	**83.0±10.1**	**95.7±4.8**	**87.5±15.0**	**93.6±4.6**	**88.4±4.4**
[2048, 1024, 512, 256]*	95.5±3.4	95.2±3.1	**95.6±2.5**	81.0±18.2	82.1±10.2	89.1±16.6	86.2±14.2	93.2±4.4	87.5±5.9
[1024, 512, 256, 128]	95.5±3.9	**95.4±3.5**	94.9±2.5	**81.5±17.2**	82.1±9.9	89.1±16.8	86.4±15.4	93.4±4.2	87.6±4.9

Group size (K)							mDSC (%) ↑	mHD95 (mm) ↓
	Pa	RAG	LAG	Du	Bl	Pr/Ut		
[4096, 2048, 1024, 512]	82.8±12.2	73.3±12.4	75.0±12.9	76.4±14.1	**87.8±16.1**	82.0±16.1	86.1±14.7	48.4±87.2
[2048, 1024, 512, 256]	83.8±11.0	75.0±11.3	74.5±12.5	**77.5±13.5**	85.7±17.3	**82.4±16.9**	**86.3±13.9**	45.0±85.1
[2048, 1024, 512, 256]*	**83.9±10.8**	**75.2±11.0**	**75.1±12.2**	77.3±13.1	85.6±17.1	82.1±16.0	86.2±14.7	48.2±83.7
[1024, 512, 256, 128]	82.8±13.1	73.1±13.5	**75.1±12.4**	76.1±15.1	87.5±16.7	82.1±17.2	86.1±15.2	**51.4±83.2**

Table 19: 5-fold cross-validation on the FeTA 2022 dataset of the varying number of groups at each encoder stage. We report mDSC (%) and mHD95 (mm) for each class and across all folds. Quantitative segmentation results are provided for external cerebrospinal fluid (eCSF), gray matter (GM), white matter (WM), ventricles (V), cerebellum (C), deep gray matter (dGM), and brainstem (B). The best results are **bolded**, while the second-best are underlined. (*) indicates a doubled number of experts in the second level.

Group size (K)	DSC (%) ↑							mDSC (%) ↑	mHD95 (mm) ↓
	eCSF	GM	WM	V	C	dGM	B		
[4096, 2048, 1024, 512]	80.5±5.7	75.1±3.1	91.1±1.5	88.1±1.4	88.2±1.7	**89.6±1.9**	86.8±2.0	87.4±1.2	**2.2±0.3**
[2048, 1024, 512, 256]	**80.6±5.9**	75.3±3.1	**91.2±1.4**	**88.2±1.3**	88.3±1.5	89.3±1.8	86.7±1.6	**87.5±1.2**	**2.2±0.4**
[2048, 1024, 512, 256]*	80.4±5.7	75.3±2.9	91.0±1.6	88.0±1.4	**88.5±1.6**	89.3±1.7	86.3±1.8	87.3±1.1	2.3±0.4
[1024, 512, 256, 128]	80.4±5.1	**75.7±2.7**	90.3±1.5	87.2±1.7	**88.3±1.7**	89.4±1.8	**87.0±1.4**	87.4±1.1	2.3±0.3

D.3 Effect of the number of slots

For a fair comparison, we keep the number of experts constant ($E_1 \in \{4, 8, 12, 16\}$ and $E_2 \in \{8, 16, 24, 32\}$) and the number of groups fixed ($K \in \{2048, 1024, 512, 256\}$).

Table 20: Quantitative segmentation results of the effect of the number of slots per expert. We report mDSC (%) and mHD95 (mm) for PDAC and Pancreas classes across datasets, including NIH, MSD, in-house, and overall. Results are provided for pancreas in NIH, PDAC, and pancreas in MSD, in-house, and overall. The best results are **bolded**, while the second-best are underlined. (*) indicates a doubled number of experts in the second level.

Slots (S)	NIH	MSD		In-house		Overall		mDSC (%) ↑	mHD95 (mm) ↓
	Pancreas	PDAC	Pancreas	PDAC	Pancreas	PDAC	Pancreas		
1	**93.0±2.1**	46.6±32.4	**88.6±6.4**	55.9±31.4	**82.2±10.9**	50.1±32.3	**88.5±7.6**	76.2±26.4	9.2±5.1
2	92.5±2.8	49.7±31.1	88.0±7.4	**61.5±26.6**	82.0±11.2	54.2±30.0	88.0±8.2	77.2±24.2	7.4±4.5
4	92.9±2.2	**51.2±30.5**	88.4±7.0	60.8±26.8	81.7±11.2	**54.8±29.5**	88.3±8.1	**77.5±23.8**	**6.9±3.9**
8	92.8±2.5	46.4±32.5	88.3±6.8	60.6±27.3	81.3±11.6	51.8±31.4	88.1±8.1	76.5±25.5	8.7±5.0

Table 21: Qualitative segmentation results of the effect of the number of slots per expert on AMOS-CT validation set. We report mDSC (%) and mHD95 (mm) for each organ. Organs include Spleen (Sp), Right Kidney (RK), Left Kidney (LK), Gallbladder (GB), Esophagus (Es), Liver (Li), Stomach (St), Aorta (Ao), Inferior Vena Cava (IVC), Pancreas (Pa), Right Adrenal Gland (RAG), Left Adrenal Gland (LAG), Duodenum (Du), Bladder (Bl), and Prostate/Uterus (Pr/Ut). The best results are **bolded**, while the second-best are underlined.

Slots (S)	DSC (%) ↑								
	Sp	RK	LK	GB	Es	Li	St	Ao	IVC
1	95.0±5.0	93.0±8.1	95.1±3.2	**80.6±22.0**	81.4±11.9	95.7±5.8	87.4±15.1	93.5±4.9	88.8±4.3
2	**95.5±3.9**	95.3±3.4	95.1±4.5	80.3±19.7	82.3±9.9	91.3±15.9	87.7±15.9	**93.7±4.4**	**88.9±4.1**
4	**95.5±4.1**	**96.0±1.5**	**95.5±2.8**	79.7±21.1	**82.4±10.1**	**96.1±3.8**	**87.8±15.1**	93.4±4.2	88.7±4.2
8	95.3±4.3	95.8±1.6	95.3±3.0	79.5±22.2	82.3±10.0	**96.1±4.0**	87.6±15.3	93.3±4.6	88.7±4.2

Slots (S)							mDSC (%) ↑	mHD95 (mm) ↓
	Pa	RAG	LAG	Du	Bl	Pr/Ut		
1	**84.4±10.8**	75.0±12.0	74.8±12.2	77.8±12.7	83.8±19.0	**82.4±15.1**	85.9±14.0	54.4±25.3
2	84.0±10.9	74.6±11.2	74.5±14.3	**78.6±12.8**	86.2±16.3	81.3±16.6	86.0±14.0	52.4±26.5
4	82.4±12.1	**75.2±11.2**	**75.5±10.2**	76.3±14.1	**87.6±14.6**	81.7±15.2	**86.3±13.9**	**45.0±23.1**
8	82.2±12.5	74.9±11.8	75.4±11.6	76.2±14.3	**87.6±15.7**	81.6±16.0	86.1±13.9	51.2±25.2

Table 22: 5-fold cross-validation on the FeTA 2022 dataset of the varying number of slots per experts at each encoder stage. We report mDSC (%) and mHD95 (mm) for each class and across all folds. Quantitative segmentation results are provided for external cerebrospinal fluid (eCSF), gray matter (GM), white matter (WM), ventricles (V), cerebellum (C), deep gray matter (dGM), and brainstem (B). The best results are **bolded**, while the second-best are underlined.

Slots (S)	DSC (%) ↑							mDSC (%) ↑	mHD95 (mm) ↓
	eCSF	GM	WM	V	C	dGM	B		
1	80.6±5.6	75.1±3.2	91.0±1.7	87.9±1.7	88.4±1.4	89.4±1.9	86.4±1.7	87.3±1.1	2.2±0.4
2	**80.7±5.6**	**75.3±3.1**	91.1±1.6	87.9±1.7	88.5±1.5	**89.6±1.8**	86.4±1.8	87.4±1.2	2.4±0.5
4	80.4±5.8	**75.3±3.1**	**91.2±1.5**	**88.2±1.3**	**88.8±1.6**	89.4±1.8	**86.6±1.5**	**87.5±1.2**	**2.1±0.2**
8	80.5±5.6	75.3±3.0	91.1±1.5	88.1±1.3	88.5±1.2	89.5±1.7	88.6±1.7	87.4±1.2	2.1±0.3

D.4 Impact of the HoME layer

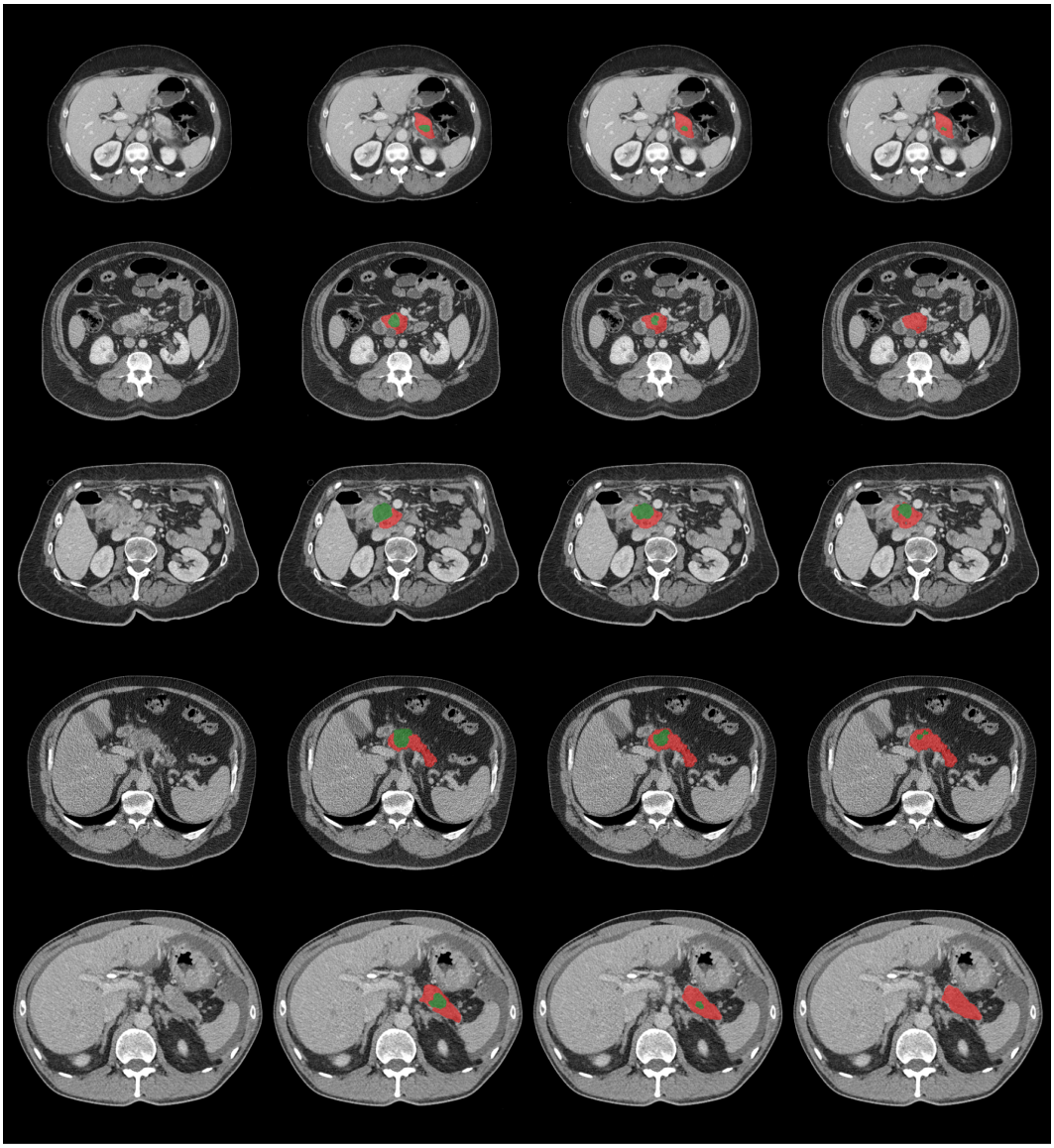

Figure 4: Qualitative comparison of Mamba-HoME and SegMamba on abdominal CT scans from PANORAMA test set. The images highlight the impact of the HoME layer added to the baseline SegMamba model, with green (PDAC) and red (pancreas) annotations indicating segmentation differences. Mamba-HoME demonstrates robustness and improved accuracy in detecting both small and large anatomical structures, like tumors, compared to SegMamba alone. Please note that we show Mamba-HoME results trained from scratch.

E Qualitative results

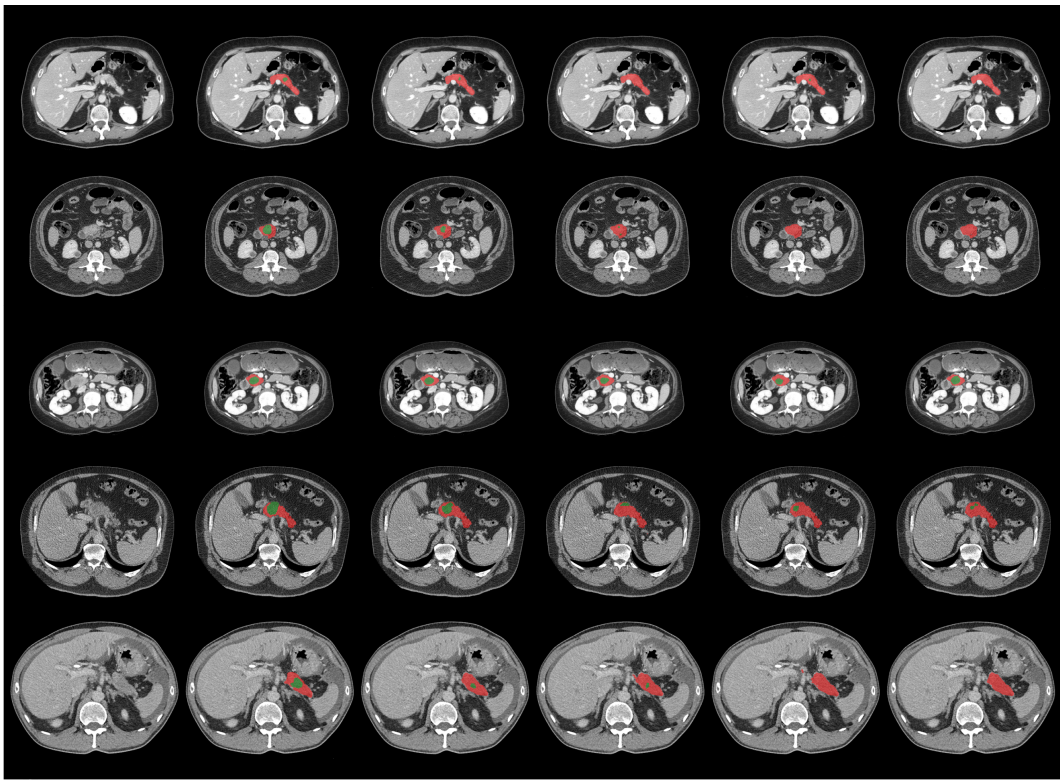

Figure 5: Qualitative segmentation results for PDAC (green) and the pancreas (red) provided by Mamba-HoME and the next three top-performing methods. The first three rows display cases from the MSD Pancreas dataset, while the last two rows show cases from the in-house dataset. Please note, we show Mamba-HoME results trained from scratch.

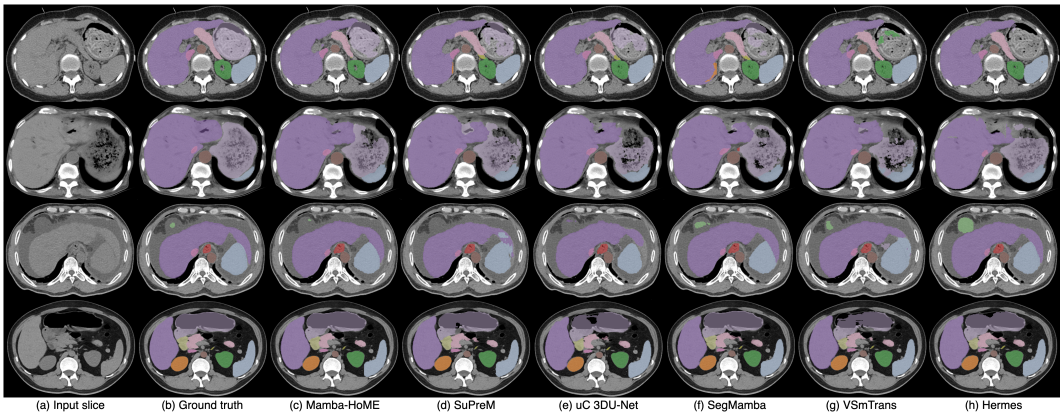

Figure 6: Qualitative segmentation results on the AMOS-CT validation set for Mamba-HoME (trained from scratch), SuPreM, uC 3DU-Net, SegMamba, VSmTrans, and Hermes. All models are trained on both the AMOS-CT and AMOS-MRI training datasets.

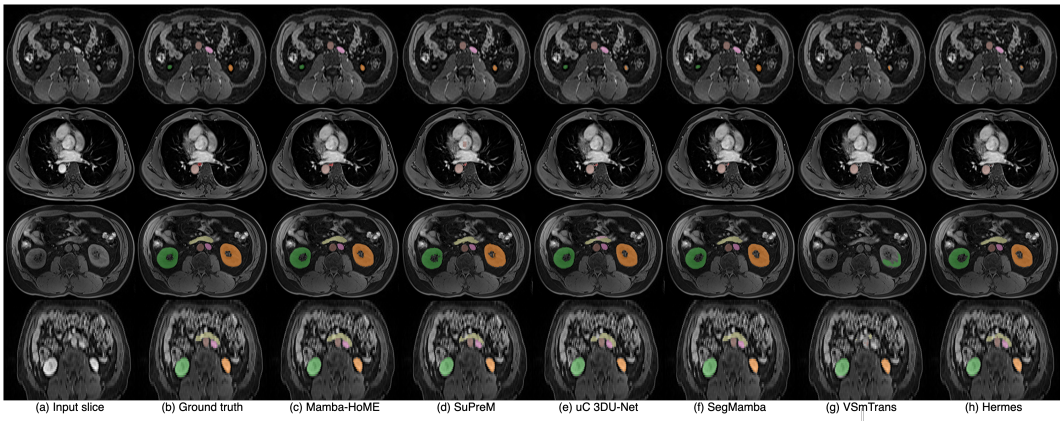

Figure 7: Qualitative comparison of segmentation performance on the AMOS-MRI validation set for six methods: Mamba-HoME (trained from scratch), SuPreM, uC 3DU-Net, SegMamba, VSmTrans, and Hermes. The models are trained on both training dataset of AMOS-CT and AMOS-MRI.

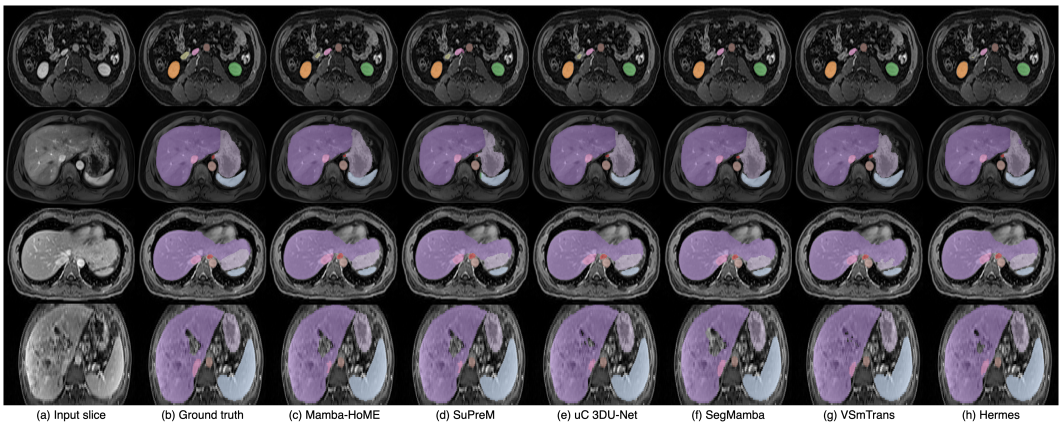

Figure 8: Qualitative segmentation results on the AMOS-MRI validation set for Mamba-HoME, SuPreM, uC 3DU-Net, SegMamba, VSmTrans, and Hermes. Models are first pre-trained on the AMOS-CT scans and subsequently fine-tuned on the AMOS-MRI training data.

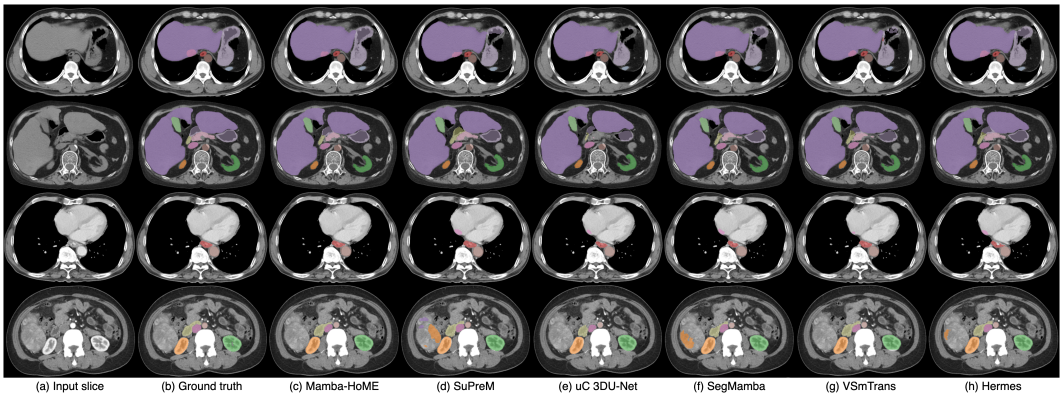

Figure 9: Qualitative segmentation results on the AMOS-CT validation set for Mamba-HoME, SuPreM, uC 3DU-Net, SegMamba, VSmTrans, and Hermes. Each model is trained only on the AMOS-CT training scans.

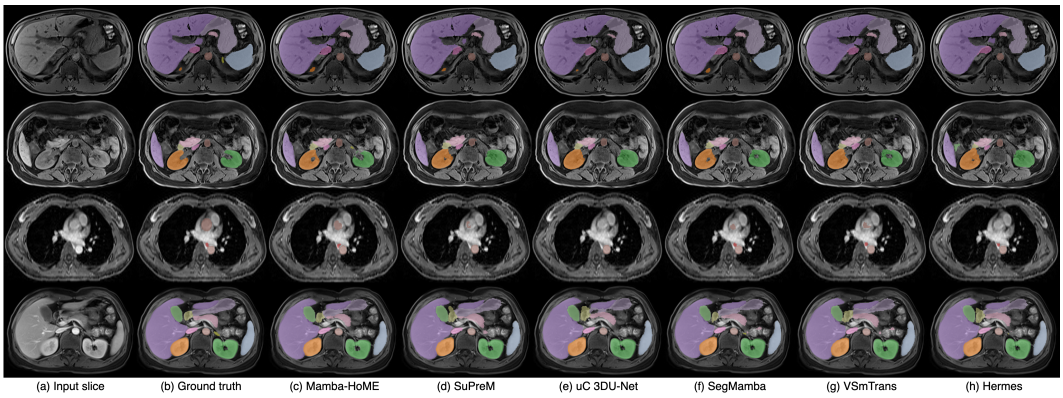

Figure 10: Qualitative segmentation results on the AMOS-MRI validation set for Mamba-HoME, SuPreM, uC 3DU-Net, SegMamba, VSmTrans, and Hermes. These models are trained solely on the AMOS-MRI training set.

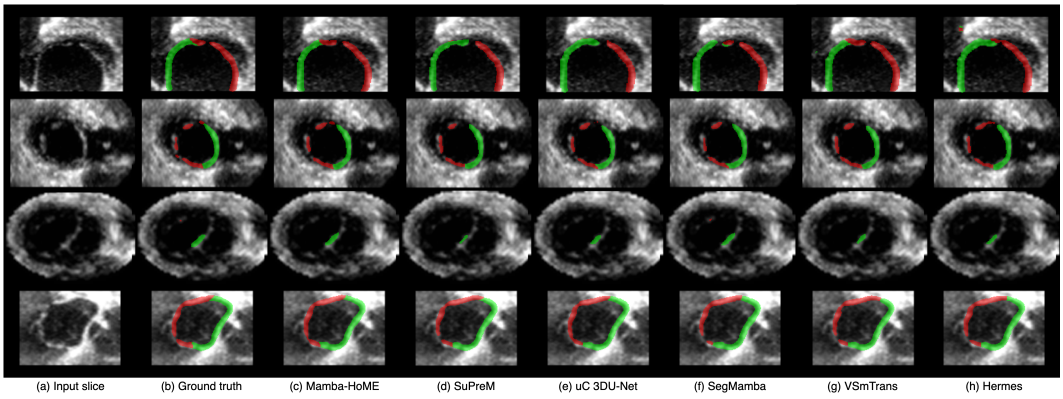

Figure 11: Qualitative segmentation results of Mamba-HoME, SuPreM, uC 3DU-Net, SegMamba, VSmTrans, and Hermes on the MVSeg test set.

F Generalizability analysis

Figure 12 provides an overview of the datasets utilized in the Mamba-HoME framework, encompassing CT, MRI, fetal MRI, and 3D ultrasound modalities. The figure presents voxel-wise ground truth labels across cross-modal and cross-anatomical domains. The first and third columns, as well as the second and fourth columns, display independent input slices and their corresponding ground truth segmentations for two representative cases. The framework, initially developed using CT and MRI scans, demonstrates consistent segmentation of abdominal organs such as the liver, spleen, and kidneys. The inclusion of fetal MRI and 3D ultrasound for fine-tuning further highlights the model's capacity to generalize across modalities with distinct anatomical features and imaging characteristics. This cross-modal and cross-anatomical representation emphasizes the versatility of the Mamba-HoME network in capturing variability across both imaging techniques and anatomical structures.

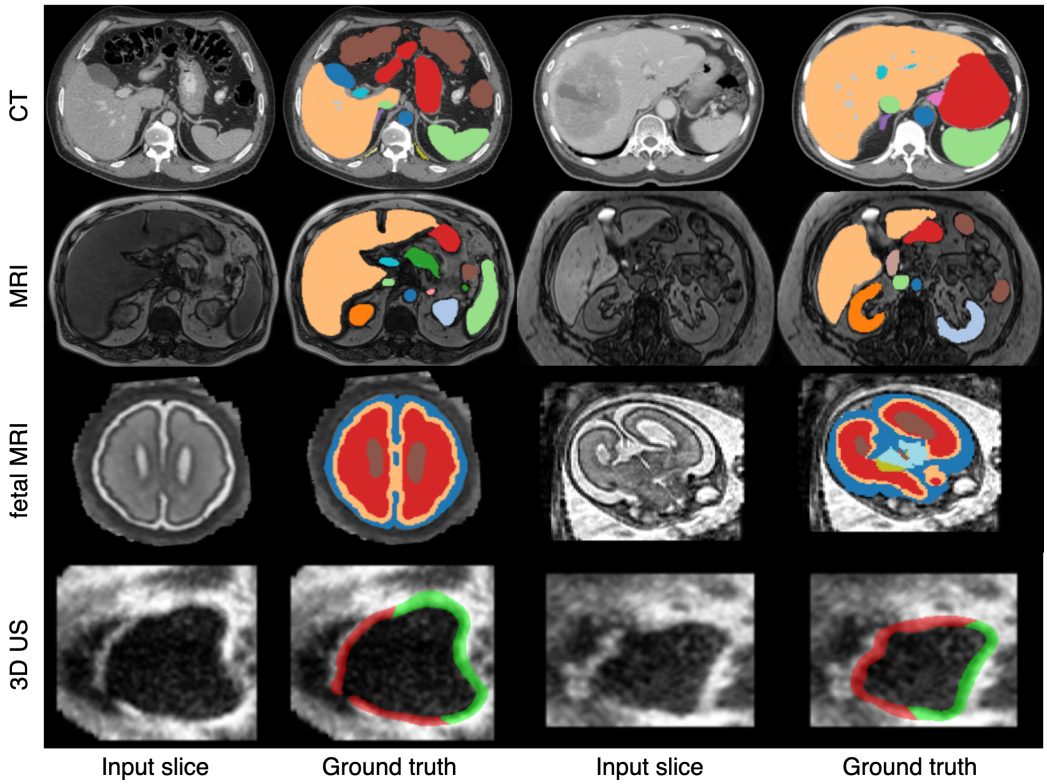

Figure 12: This figure provides an overview of the dataset used for pre-training, including CT scans (first row), MRI scans (second row), and fetal MRI scans (third row), with 3D ultrasound (fourth row) specifically used for fine-tuning. The first and third columns, as well as the second and fourth columns, display independent input slices and corresponding ground truth for two cases. Although fetal MRI and 3D ultrasound fall under the broader category of CT and MRI-based exams, this comparison highlights the distinct differences in feature representation between both CT, MRI and fetal MRI with 3D ultrasound, showcasing variations in anatomical detail and imaging characteristics across these modalities.

We provide detailed quantitative segmentation results from the generalizability analysis. In this experiment, we train each network using the following protocols: (1) we train the model solely on the AMOS-CT training scans and evaluate it on the AMOS-CT validation set (see Table 23); (2) we train a model solely on the AMOS-MRI training scans and evaluate it on the AMOS-MRI validation set (see Table 24); (3) we train the model on the AMOS-MRI training scans with pre-training on the AMOS-CT training scans, and evaluate it on the AMOS-MRI validation set (see Table 25); and (4) we train the model on both the AMOS-CT and AMOS-MRI training scans, and evaluate it on both the AMOS-CT and AMOS-MRI validation sets (see Table 26).

Table 23: Qualitative segmentation results on AMOS-CT validation set. We report mDSC (%) and HD95 (mm) for each organ. Organs include Spleen (Sp), Right Kidney (RK), Left Kidney (LK), Gallbladder (GB), Esophagus (Es), Liver (Li), Stomach (St), Aorta (Ao), Inferior Vena Cava (IVC), Pancreas (Pa), Right Adrenal Gland (RAG), Left Adrenal Gland (LAG), Duodenum (Du), Bladder (Bl), and Prostate/Uterus (Pr/Ut). The best results are **bolded**, while the second-best are underlined. (*) indicates pre-trained model.

Method	DSC (%) ↑								
	Sp	RK	LK	GB	Es	Li	St	Ao	IVC
Hermes	95.7±2.3	94.5±4.4	95.3±2.2	80.9±19.5	81.1±11.0	93.9±9.5	88.0±14.6	92.9±4.1	87.8±5.6
Swin SMT	95.4±5.4	**95.6±1.8**	95.1±2.5	76.3±24.0	81.3±10.1	92.2±12.8	87.4±14.3	93.7±3.5	88.6±4.4
VSmTrans	95.6±2.5	94.7±3.1	94.5±3.2	82.7±17.1	76.4±12.0	95.2±4.1	86.9±13.6	92.9±4.8	83.1±10.8
SegMamba	95.7±3.3	95.5±1.9	95.2±2.9	81.7±18.8	81.1±10.3	94.9±6.2	87.7±15.1	93.2±4.1	88.8±3.8
uC 3DU-Net	93.0±8.8	94.6±3.2	94.3±5.2	73.6±24.3	78.8±12.3	93.4±9.6	82.5±17.2	92.4±4.3	85.5±6.3
Swin UNETR	95.2±4.8	95.1±2.6	94.5±4.3	72.2±27.8	80.4±12.2	90.2±15.2	81.0±22.6	93.2±4.1	87.9±5.1
SuPreM (*)	95.6±2.7	95.0±1.7	94.4±5.0	**86.2±12.0**	79.4±9.7	**97.1±1.2**	91.2±9.4	93.0±3.2	88.7±4.0
Mamba-HoME	**96.0±2.5**	95.0±4.2	94.4±4.1	81.7±19.9	82.0±10.2	94.7±7.9	90.2±13.1	**93.8±3.0**	**88.9±4.0**
Mamba-HoME (*)	**96.0±2.5**	95.4±1.8	**95.4±2.2**	85.3±15.5	**82.1±9.5**	96.8±1.7	**91.3±10.2**	93.3±4.0	88.6±4.3

Method							mDSC (%) ↑	mHD95 (mm) ↓
	Pa	RAG	LAG	Du	Bl	Pr/Ut		
Hermes	82.5±12.2	73.7±13.1	71.8±15.4	77.7±14.4	83.3±19.9	79.3±17.3	85.3±14.7	40.5±87.2
Swin SMT	84.1±10.1	74.4±12.2	73.4±13.3	78.3±13.5	86.6±17.1	82.6±14.8	85.7±14.3	91.4±153.2
VSmTrans	84.0±10.0	72.9±11.2	**74.7±12.0**	77.8±12.7	86.8±15.6	81.6±16.2	85.3±13.4	178.7±190.5
SegMamba	84.0±10.0	**74.7±11.4**	74.0±12.7	77.3±14.1	83.9±17.5	82.6±12.5	86.0±13.3	98.9±138.9
uC 3DU-Net	75.4±16.0	73.3±11.3	71.1±14.6	69.8±17.0	83.0±17.1	79.1±15.6	82.7±16.0	47.3±91.7
Swin UNETR	81.6±13.1	74.4±11.5	74.5±11.5	78.1±13.6	85.0±16.5	80.8±13.8	84.3±15.7	94.7±156.2
SuPreM (*)	**85.3±8.8**	69.3±10.0	68.8±11.5	**81.3±8.6**	86.0±15.1	78.9±15.4	86.0±12.6	66.0±100.2
Mamba-HoME	84.0±10.8	74.4±11.3	74.1±13.1	78.4±13.5	83.8±18.7	**83.2±14.9**	86.3±13.5	45.0±85.1
Mamba-HoME (*)	84.7±10.1	74.0±11.7	74.6±11.8	80.8±11.4	**88.0±13.8**	82.8±15.5	**87.3±12.1**	**32.0±64.2**

Table 24: Qualitative segmentation results on AMOS-MRI validation set. We train each model solely on MRI data. We report mDSC (%) and mHD95 (mm) for each organ. Organs include Spleen (Sp), Right Kidney (RK), Left Kidney (LK), Gallbladder (GB), Esophagus (Es), Liver (Li), Stomach (St), Aorta (Ao), Inferior Vena Cava (IVC), Pancreas (Pa), Right Adrenal Gland (RAG), Left Adrenal Gland (LAG), Duodenum (Du), Bladder (Bl), and Prostate/Uterus (Pr/Ut). The best results are **bolded**, while the second-best are underlined. (*) indicates pre-trained model.

Method	DSC (%) ↑								
	Sp	RK	LK	GB	Es	Li	St	Ao	IVC
Hermes	95.5±2.3	95.4±1.1	95.3±1.1	68.0±22.1	**72.3±11.4**	96.7±1.5	83.4±13.8	90.2±5.4	87.2±4.0
Swin SMT	92.8±4.0	94.2±2.2	93.7±2.4	56.5±26.9	0.0±0.0	95.8±2.4	81.1±15.7	88.6±4.9	83.2±6.3
VSmTrans	94.8±2.0	94.3±3.6	94.9±2.2	65.3±24.9	69.7±9.2	96.6±1.6	83.8±16.2	90.6±4.1	85.5±6.1
SegMamba	95.6±1.9	95.1±1.7	95.3±1.4	64.3±27.2	70.2±11.4	96.9±1.1	87.1±7.7	**91.2±4.1**	86.6±6.0
uC 3DU-Net	91.6±5.6	92.6±2.9	93.4±2.4	71.0±25.5	67.1±12.1	95.0±4.1	83.2±13.6	89.0±6.1	83.4±7.3
Swin UNETR	93.5±4.2	94.5±2.5	94.7±1.9	70.3±22.5	68.2±9.4	96.3±2.6	84.4±14.1	90.2±4.5	86.0±4.5
SuPreM (*)	93.6±4.9	94.6±1.9	94.0±2.1	71.2±23.7	67.3±10.3	95.9±3.8	83.7±11.3	89.3±5.2	83.8±6.2
Mamba-HoME	95.2±2.1	94.6±2.1	94.6±2.4	67.6±27.5	70.6±9.4	96.7±1.5	85.0±10.7	90.0±5.1	**87.6±5.0**
Mamba-HoME (*)	**96.1±1.3**	**95.5±1.1**	**95.7±0.9**	**75.3±22.4**	**72.3±8.9**	**97.5±0.5**	**89.1±6.7**	90.4±5.4	86.5±5.4

Method							mDSC (%) ↑	mHD95 (mm) ↓
	Pa	RAG	LAG	Du	Bl	Pr/Ut		
Hermes	79.2±19.5	61.9±10.8	60.8±16.9	63.1±14.3	NA	NA	80.7±17.9	**13.9±33.5**
Swin SMT	79.4±12.5	0.0±0.0	0.0±0.0	56.3±13.0	NA	NA	63.2±38.2	45.1±72.7
VSmTrans	80.3±16.7	52.4±17.4	0.0±0.0	61.5±16.4	NA	NA	74.6±28.3	30.6±53.8
SegMamba	80.2±18.3	56.3±18.4	59.8±13.7	63.7±13.9	NA	NA	80.2±19.2	33.1±52.2
uC 3DU-Net	79.4±16.1	0.0±0.0	0.0±0.0	59.8±14.1	NA	NA	68.6±35.2	16.3±28.5
Swin UNETR	80.5±15.3	54.5±12.3	0.0±0.0	61.5±16.0	NA	NA	75.0±27.5	30.7±51.4
SuPreM (*)	79.8±15.0	0.0±0.0	0.0±0.0	60.6±13.0	NA	NA	70.3±33.4	52.1±69.8
Mamba-HoME	81.6±15.6	**62.7±12.1**	60.5±11.6	65.8±13.8	NA	NA	81.0±17.4	32.5±52.3
Mamba-HoME (*)	**85.3±9.8**	58.2±15.7	**62.1±9.7**	**65.9±13.8**	NA	NA	**82.3±16.7**	19.5±31.2

Table 25: Qualitative segmentation results on AMOS-MRI validation set, pre-trained on AMOS-CT dataset (CT → MRI). We report mDSC (%) and mHD95 (mm) for each organ. Organs include Spleen (Sp), Right Kidney (RK), Left Kidney (LK), Gallbladder (GB), Esophagus (Es), Liver (Li), Stomach (St), Aorta (Ao), Inferior Vena Cava (IVC), Pancreas (Pa), Right Adrenal Gland (RAG), Left Adrenal Gland (LAG), Duodenum (Du), Bladder (Bl), and Prostate/Uterus (Pr/Ut). The best results are **bolded**, while the second-best are underlined. (*) indicates pre-trained model.

Method	DSC (%) ↑								
	Sp	RK	LK	GB	Es	Li	St	Ao	IVC
Hermes	**96.6±1.2**	**95.8±0.8**	96.0±0.7	77.7±27.7	**78.3±6.3**	97.7±0.7	90.4±8.5	**92.5±3.8**	89.9±2.9
Swin SMT	96.1±1.2	95.4±0.9	95.7±0.7	75.7±19.3	75.2±7.4	97.3±0.7	88.2±10.3	91.7±3.5	89.1±3.2
VSmTrans	96.2±1.3	**95.8±0.8**	95.7±0.7	79.9±18.3	74.1±7.8	87.5±0.6	90.5±4.9	91.2±4.2	89.5±2.9
SegMamba	94.7±2.9	95.0±1.8	95.0±1.4	67.4±22.7	69.9±9.7	96.3±2.3	84.3±13.0	90.5±3.0	84.7±6.6
uC 3DU-Net	96.1±1.4	**95.8±0.8**	95.8±0.7	**80.1±21.2**	76.8±8.2	97.3±1.0	88.5±10.1	91.8±3.9	89.1±3.1
Swin UNETR	96.1±1.1	95.6±0.7	95.8±0.6	78.9±15.5	72.4±7.6	97.2±0.8	89.2±6.5	91.0±4.1	88.9±3.0
SuPreM (*)	96.1±1.9	95.3±2.0	95.8±0.7	76.7±22.5	73.9±7.4	97.5±0.7	88.9±9.1	91.8±3.7	88.8±3.3
Mamba-HoME	96.4±1.4	95.5±0.9	95.9±0.6	77.3±22.2	76.2±8.7	97.5±0.7	90.9±5.0	91.4±5.5	89.4±2.9
Mamba-HoME (*)	96.5±1.3	**95.8±0.9**	**96.1±0.7**	79.5±22.7	77.0±8.1	**97.8±0.6**	**91.2±5.1**	92.4±3.0	**90.4±2.8**

Method							mDSC (%) ↑	mHD95 (mm) ↓
	Pa	RAG	LAG	Du	Bl	Pr/Ut		
Hermes	86.0±12.9	65.2±9.9	63.5±16.0	72.0±13.3	NA	NA	84.8±16.0	**11.9±20.1**
Swin SMT	82.3±18.4	62.9±13.5	65.2±12.6	65.9±13.8	NA	NA	83.2±16.1	21.0±41.7
VSmTrans	85.4±12.6	64.2±9.6	66.8±9.3	70.8±11.7	NA	NA	84.5±14.2	30.1±55.3
SegMamba	80.8±14.6	57.0±15.3	50.5±17.1	60.6±13.7	NA	NA	79.0±19.4	14.7±26.6
uC 3DU-Net	84.3±16.0	63.8±10.8	68.4±9.0	71.0±13.2	NA	NA	84.5±14.8	15.4±25.8
Swin UNETR	85.1±13.8	**65.9±9.1**	**69.3±8.2**	68.6±15.4	NA	NA	84.2±14.1	28.4±57.7
SuPreM (*)	85.2±13.4	63.5±13.9	66.5±15.4	71.0±13.1	NA	NA	83.9±15.7	14.8±26.9
Mamba-HoME	**86.4±12.1**	64.5±9.9	66.7±10.5	**72.2±12.8**	NA	NA	84.8±14.7	16.0±37.3
Mamba-HoME (*)	85.4±17.1	63.9±11.1	66.5±12.2	72.0±14.2	NA	NA	**85.0±15.5**	13.3±30.1

Table 26: Qualitative segmentation results on AMOS-CT + AMOS-MRI validation sets. We train each model on both CT and MRI training sets and evaluate on both CT and MRI validation sets. We report mDSC (%) and mHD95 (mm) for each organ. Organs include Spleen (Sp), Right Kidney (RK), Left Kidney (LK), Gallbladder (GB), Esophagus (Es), Liver (Li), Stomach (St), Aorta (Ao), Inferior Vena Cava (IVC), Pancreas (Pa), Right Adrenal Gland (RAG), Left Adrenal Gland (LAG), Duodenum (Du), Bladder (Bl), and Prostate/Uterus (Pr/Ut). The best results are **bolded**, while the second-best are underlined. (*) indicates pre-trained model.

Method	DSC (%) ↑								
	Sp	RK	LK	GB	Es	Li	St	Ao	IVC
Hermes	94.5±2.3	93.7±4.1	93.3±2.0	77.8±20.5	77.6±11.5	93.8±8.8	87.2±14.6	92.5±4.4	87.7±5.4
Swin SMT	91.1±12.8	94.0±4.6	94.2±4.1	75.1±23.8	76.7±13.0	94.0±8.4	79.2±21.4	90.8±6.2	85.6±12.2
VSmTrans	85.0±27.5	87.1±22.7	89.1±18.5	71.7±30.3	73.9±21.2	86.3±25.1	74.5±29.5	87.5±17.2	79.2±24.0
SegMamba	94.9±6.7	94.0±5.9	93.5±7.2	76.7±23.5	79.7±11.8	94.5±8.5	89.1±11.7	**93.1±4.3**	**88.3±4.4**
uC 3DU-Net	95.0±3.7	94.5±3.3	93.8±7.5	76.3±23.9	78.8±12.5	95.3±6.2	86.9±14.0	92.7±4.0	87.6±4.6
Swin UNETR	92.1±12.5	94.8±3.3	94.6±3.8	78.1±22.1	76.8±13.5	95.1±6.4	86.5±13.8	91.8±5.2	84.7±10.7
SuPreM (*)	92.8±11.4	92.6±13.1	94.1±5.5	78.5±22.7	78.1±13.1	93.2±13.8	84.4±18.9	92.3±5.1	84.5±14.0
Mamba-HoME	**95.2±4.5**	95.1±2.1	94.6±4.2	79.3±21.4	79.9±10.8	95.4±7.4	89.1±11.7	92.8±4.4	88.2±4.3
Mamba-HoME (*)	95.0±6.9	**95.3±2.6**	**95.6±1.8**	**82.7±19.5**	**80.0±12.6**	**96.3±2.8**	**90.4±10.8**	93.0±4.5	87.8±6.2

Method							mDSC (%) ↑	mHD95 (mm) ↓
	Pa	RAG	LAG	Du	Bl	Pr/Ut		
Hermes	82.0±13.8	71.7±13.5	70.0±16.2	72.2±15.4	83.3±20.0	79.3±17.3	82.9±15.3	36.1±80.7
Swin SMT	80.3±13.4	68.1±16.5	70.3±13.0	68.8±17.6	70.3±23.3	76.8±15.7	81.2±17.6	98.9±148.3
VSmTrans	74.5±24.0	64.3±25.9	67.2±21.7	67.3±25.8	82.6±19.3	80.8±12.0	78.0±24.9	102.1±131.8
SegMamba	83.1±12.0	68.5±16.2	72.5±13.7	76.0±14.8	83.4±19.8	82.3±14.9	84.7±15.3	64.2±124.9
uC 3DU-Net	81.9±12.4	71.7±14.0	70.7±15.4	74.0±15.8	83.2±19.4	76.7±20.3	84.1±15.8	34.9±78.7
Swin UNETR	82.8±12.8	69.1±13.9	71.0±15.0	74.6±17.3	83.1±18.8	81.5±13.7	83.8±15.7	84.7±145.1
SuPreM (*)	82.3±12.5	67.8±19.4	71.9±13.9	74.2±16.5	84.4±18.2	81.2±13.4	83.5±16.7	48.2±86.3
Mamba-HoME	82.7±12.2	71.9±13.5	72.7±13.4	75.5±13.7	82.7±20.8	80.5±16.7	85.1±14.6	54.8±75.2
Mamba-HoME (*)	**84.5±11.7**	**72.1±12.9**	**72.9±13.7**	**77.7±14.3**	**89.0±14.5**	**83.2±15.8**	**86.4±13.8**	**27.7±61.2**

G Impact statement

This work introduces Hierarchical Soft Mixture-of-Experts (HoME), a novel architecture for efficient and accurate 3D medical image segmentation. By integrating a two-level mixture-of-experts routing mechanism with Mamba-based Selective State Space Models, our method significantly advances long-context modeling for volumetric data. HoME is designed to address key challenges in medical imaging, namely, modeling local-to-global spatial hierarchies, handling modality diversity (CT, MRI, US), and achieving scalability for high-resolution 3D inputs. Our proposed Mamba-HoME architecture demonstrates strong generalization and outperforms state-of-the-art models across public and in-house datasets, while being memory and compute efficient.

Beyond medical imaging, the architectural principles introduced, specifically the hierarchical token routing and the integration of local and global context processing, are applicable to other domains dealing with structured, hierarchical data under resource constraints. These include scientific computing, robotics, and spatiotemporal analysis in environmental or geospatial datasets.

Ethically, this work supports equitable healthcare by enabling accurate segmentation with reduced computational requirements, which is crucial for deployment in low-resource settings. We use publicly available datasets and provide open-source code to ensure reproducibility and accessibility for the broader community. No personally identifiable information or sensitive patient data is used. Future extensions could include further robustness to distributional shifts in medical data and broader clinical evaluation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In both abstract and introduction, we accurately reflect the paper's contributions and scope. Specifically, we introduce Hierarchical Soft Mixture-of-Experts (HoME), a two-level, token-routing MoE layer for efficient capture of local-to-global pattern hierarchies. Additionally, we design a unified architectural block that integrates Mamba's SSMs with HoME. We embed the above novel solutions into a multi-stage U-shaped architecture, called Mamba-HoME, designed for 3D medical image segmentation.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In conclusion, we provided a Limitations section, where we list the main limitations of the paper. First, scalability to large-scale medical datasets and self-supervised pre-training is a promising direction for future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper focuses on empirical results, and we do not provide any theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper details the experimental setups, dataset descriptions, hyperparameters, and Mamba-HoME architecture. The code and four datasets used for training and evaluation are publicly available in the supplementary material, and an in-house dataset used for testing is available upon request.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code and four datasets used for evaluation are publicly available. An in-house dataset used for generalizability analysis is available upon request.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides descriptions of training and testing, including data splits, hyperparameter choices, and the optimizer type.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We use the Wilcoxon signed-rank test to evaluate the statistical significance of performance differences between our proposed method, Mamba-HoME, and other state-of-the-art approaches.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper details the computational resources used for training. Specifically, we used Python 3.11, PyTorch 2.4, and MONAI 1.4.0. For hardware, we employed an NVIDIA DGX system equipped with 8 $\times$ NVIDIA H100 80 GB GPUs. We used all eight GPUs for pre-training, while training, fine-tuning, and evaluation were performed on a single NVIDIA H100 80 GB GPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research presented in this paper conforms in every respect to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper describes the efficiency and accuracy of 3D medical image segmentation, which can enhance medical diagnostic precision and patient outcomes.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not present any associated risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In this paper, we acknowledge the original authors of all utilized assets, including code, datasets, and models, by citing them appropriately within the text.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The proposed Hierarchical Soft Mixture-of-Experts (HoME) and its integration into Mamba-HoME are fully described in the methodology section, including architecture, implementation, and datasets used for validation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: We have obtained IRB approval for our in-house data used as part of the test set. Due to the double-blind review process, we will release the IRB number upon acceptance.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We do not use LLM for the core methodology, scientific rigorousness, or originality of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.