AAMAMBA: NOT ALL VOXELS DESERVE EQUAL TOKENS IN MAMBA-BASED 3D MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical volumes are highly nonuniform. Most voxels belong to homogeneous background or simple tissue, while the evidence needed for accurate segmentation concentrates at lesion margins, organ interfaces, and small irregular structures. State-space models such as Mamba process long 3D sequences efficiently, yet existing Mamba-based segmentation models still build tokens on uniform grids and serialize them in fixed spatial orders. We identify this anatomy-agnostic volume-to-sequence construction as a central bottleneck of 3D state-space segmentation. Uniform tokenization gives background the same representational budget as lesions and boundaries, and naive scan orders can place neighboring voxels far apart once the volume is flattened. We propose AAMamba, an anatomy-aware Mamba framework in which token allocation, sequence serialization, and dense reconstruction are designed jointly. AAMamba estimates a structural density field from intensity deviation, boundary strength, and curvature response, and uses it to assign denser tokens to anatomically complex regions while still covering the whole volume. The adaptive tokens are then ordered by ROI-guided bidirectional Hilbert scanning, which keeps spatial neighbors close in the sequence. ROI guidance is applied only during training to emphasize foreground structures, so inference requires no ROI input. Because adaptive sampling can disturb voxel-level correspondence, a coordinate-aware decoder fuses the adaptive semantic features with grid-aligned spatial references to reduce spatial drift in the dense prediction. AAMamba achieves 84.2% mean Dice on BraTS, 91.85% DSC on ACDC, and 84.26% average Dice on Synapse. The largest gains appear on small or morphologically complex structures, including the enhancing tumor, myocardium, gallbladder, and pancreas, which suggests that anatomy-aware sequence construction is an effective principle for 3D medical state-space models.

1 INTRODUCTION

Accurate 3D medical image segmentation requires both global volumetric context and precise local geometry. Convolutional encoder–decoder models such as U-Net, 3D U-Net, V-Net, and nnU-Net remain strong baselines because they preserve spatial structure through hierarchical features and skip connections (Ronneberger et al., 2015; Çiçek et al., 2016; Milletari et al., 2016; Isensee et al., 2021). However, anatomical information is highly nonuniform: large regions may be homogeneous, while clinically important evidence is concentrated around lesion margins, organ interfaces, low-contrast boundaries, and small abnormal structures. This suggests that a segmentation model should not allocate the same representational effort everywhere. Transformer-based models improve long-range reasoning (Dosovitskiy et al., 2020; Chen et al., 2021; Hatamizadeh et al., 2022b;a; Zhou et al., 2021), but volumetric attention is expensive because token count grows rapidly with spatial resolution. State-space models such as S4 and Mamba offer a more efficient alternative (Gu et al., 2021; Gu & Dao, 2024), and recent vision and medical Mamba models have shown promising results (Zhu et al., 2024; Liu et al., 2024; Ma et al., 2024; Ruan et al., 2024; Lu et al., 2025). Yet applying Mamba to 3D segmentation raises a fundamental question: *how should a structured 3D volume be converted into a 1D sequence?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

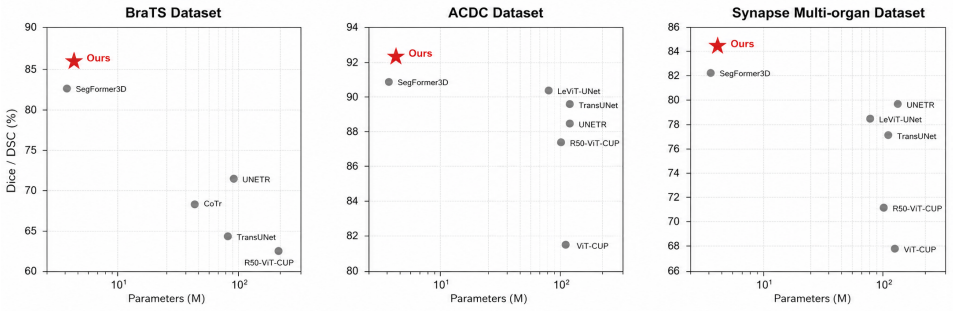

Figure 1: Accuracy–parameter comparison on BraTS, ACDC, and Synapse. AAMamba achieves competitive segmentation accuracy with a small parameter budget.

We argue that **volume-to-sequence construction is a central bottleneck in 3D state-space segmentation**. Existing Mamba-style vision models typically rely on fixed raster, axial, Cartesian, or other predefined scan orders (Zhu et al., 2024; Liu et al., 2024; Ma et al., 2024; Liu et al., 2025). This creates two mismatches. First, uniform tokenization assigns equal capacity to homogeneous background, organ interiors, complex boundaries, and small lesions, despite their very different segmentation difficulty. Second, fixed flattening can distort 3D locality by placing nearby voxels far apart in the sequence and unrelated regions next to each other across scan boundaries. Related adaptive representation methods, including dynamic token pruning, token selection, TokenLearner, and deformable convolution (Rao et al., 2021a; Xu et al., 2022; Rao et al., 2021b; Ryoo et al., 2021; Dai et al., 2017), reduce redundancy or improve feature alignment, but do not jointly address *where representation should be allocated, how it should be serialized, and how dense spatial correspondence should be recovered*.

To address this, we propose **AAMamba**, an anatomy-aware Mamba framework that treats sequence construction as part of the segmentation model itself. AAMamba first estimates a structural density field from intensity deviation, gradient magnitude, and curvature response, allowing denser representation in anatomically complex regions while preserving full-volume coverage. The resulting adaptive tokens are serialized using **ROI-guided bidirectional Hilbert scanning**: Hilbert traversal preserves local neighborhoods during 3D-to-1D conversion, bidirectional processing reduces directional bias, and ROI guidance emphasizes foreground-relevant structures during training without requiring ROI input at inference. Because adaptive sampling may disturb voxel-level correspondence, AAMamba also maintains a grid-aligned stream and uses a **coordinate-aware decoder** to fuse adaptive semantic features with spatially stable references. Thus, token allocation, sequence ordering, and dense reconstruction are designed jointly within one anatomy-aware pipeline.

We evaluate AAMamba on BraTS, ACDC, and Synapse. As shown in Figure 1, with only 5.7M parameters, AAMamba achieves 84.2% mean Dice on BraTS, 91.85% DSC on ACDC, and 84.26% average Dice on Synapse. Compared with SegFormer3D, it improves performance by 2.1, 0.89, and 1.87 Dice points, respectively, with particularly strong gains on small or morphologically complex structures such as enhancing tumor, myocardium, gallbladder, and pancreas. These results support our central claim: **for compact 3D state-space segmentation, how a volume is tokenized and serialized can be as important as the sequence model itself.**

The main contributions are:

- We identify **anatomy-aware volume-to-sequence construction** as a key design problem in 3D state-space segmentation, jointly considering token allocation, scan ordering, and coordinate reconstruction.
- We propose a **content-adaptive tokenizer** that allocates higher representational density to anatomically complex regions while preserving full-volume coverage.
- We introduce **ROI-guided bidirectional Hilbert scanning** to preserve 3D locality, reduce directional bias, and exploit foreground guidance without requiring ROI input at inference.
- We design a **coordinate-aware decoder** that combines adaptive semantic features with grid-aligned spatial references for stable dense reconstruction.

- We demonstrate consistent accuracy–efficiency gains on BraTS, ACDC, and Synapse with only 5.7M parameters.

2 Related Work

2.1 3D Medical Image Segmentation

CNN-based encoder–decoder networks remain strong foundations for medical image segmentation because they preserve local spatial detail through convolutional features, hierarchical representations, and skip connections. Representative methods include U-Net, 3D U-Net, V-Net, and nnU-Net (Ronneberger et al., 2015; Çiçek et al., 2016; Milletari et al., 2016; Isensee et al., 2021). To improve long-range reasoning, transformer-based models introduce token-based global context modeling. TransUNet combines CNN features with transformer encoders (Chen et al., 2021), while UNETR, Swin UNETR, and nnFormer adapt transformer architectures to volumetric segmentation (Hatamizadeh et al., 2022b;a; Zhou et al., 2021). Other hybrid or efficient models, including CoTr, TransBTS, LeViT-UNet, MISSFormer, and SegFormer3D, explore different trade-offs between accuracy and efficiency (Xie et al., 2021; Wang et al., 2021; Xu et al., 2023; Huang et al., 2021; Perera et al., 2024). However, attention-based models remain expensive in 3D because token counts grow cubically with spatial resolution. More importantly, most methods still begin from content-agnostic volumetric grids, assigning similar representation to homogeneous background and clinically important structures. AAMamba addresses this limitation by adapting token construction according to anatomical complexity before long-range sequence modeling.

2.2 Adaptive Tokenization and Deformable Representation

Most token-based vision models follow the regular patch-token formulation of the Vision Transformer (Dosovitskiy et al., 2020). This representation is simple and efficient, but it allocates a uniform token budget across spatial regions regardless of structural complexity. Dynamic token methods reduce redundancy by pruning, selecting, or evolving tokens after uniform tokenization (Rao et al., 2021a;b; Xu et al., 2022; Sah et al., 2024; Ye et al., 2025), while TokenLearner constructs compact informative tokens for visual sequence modeling (Ryoo et al., 2021). Deformable convolution and deformable visual models further show that adaptive spatial sampling can improve feature alignment and representation flexibility (Dai et al., 2017; Liu et al., 2025). These methods motivate adaptive representation, but they do not directly solve dense 3D medical segmentation, where the model must preserve full-volume coverage, emphasize sparse anatomical structures, and reconstruct voxel-level masks in the original coordinate system. AAMamba therefore uses a content-adaptive tokenizer to allocate denser representation to boundaries, lesions, and irregular organs, while retaining a grid-aligned stream for coordinate-stable decoding.

2.3 State-Space Models and Volumetric Sequence Construction

Structured state-space models provide efficient long-range sequence modeling with more favorable scaling than quadratic self-attention (Gu et al., 2021; Gu & Dao, 2024). Vision Mamba and VMamba extend Mamba-style selective state-space modeling to visual recognition (Zhu et al., 2024; Liu et al., 2024), and recent medical Mamba models such as U-Mamba, VM-UNet, and 3D-UMamba demonstrate the potential of state-space architectures for biomedical segmentation (Ma et al., 2024; Ruan et al., 2024; Lu et al., 2025). However, applying Mamba to 3D segmentation requires converting volumetric features into one-dimensional sequences. Existing methods often rely on raster, axial, Cartesian, or other predefined scan orders, which may disrupt local 3D neighborhoods after flattening and create a mismatch between one-dimensional state propagation and three-dimensional anatomical structure. AAMamba focuses on this underexplored sequence-construction problem. It uses bidirectional Hilbert scanning to better preserve volumetric locality, training-time ROI guidance to emphasize foreground-relevant anatomy without requiring ROI input at inference, and a coordinate-aware decoder to recover dense voxel-level predictions from adaptive sequence representations.

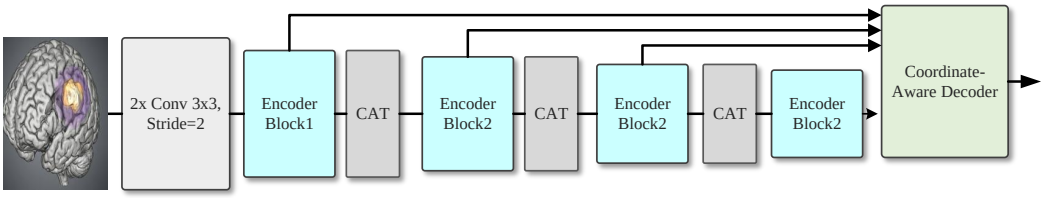

Figure 2: Overview of the AAMamba architecture. The input 3D volume is first embedded by a $2\times$ strided 3D convolution and then processed by a hierarchical encoder. Each stage consists of a content-adaptive tokenizer (CAT) followed by an ROI-guided Hilbert Mamba encoder block, where CAT reallocates representation toward anatomically complex regions and the encoder block performs locality-preserving state-space modeling with training-time ROI guidance. Multi-stage features are forwarded through skip connections to the coordinate-aware decoder, which fuses anatomy-adaptive semantic features with spatially aligned representations to produce dense voxel-level segmentation masks.

3 METHOD

3.1 OVERVIEW

AAMamba is a hierarchical encoder–decoder framework for adaptive 3D medical image segmentation. It addresses two representation problems. First, uniform tokenization gives homogeneous background and anatomically complex regions the same token budget, although lesions, boundaries, and small organs require greater modeling precision. Second, Mamba processes tokens as one-dimensional sequences, so the scan order used to serialize 3D features directly affects spatial continuity. AAMamba therefore contains three components: a content-adaptive tokenizer, an ROI-guided bidirectional Hilbert Mamba block, and a coordinate-aware decoder.

Given an encoder feature map $x_n \in \mathbb{R}^{B\times C_n\times D_n\times H_n\times W_n}$ at stage n, the tokenizer produces two aligned streams: adaptive tokens P_n for anatomy-aware context modeling and grid-aligned tokens G_n for preserving an undistorted spatial reference. The adaptive tokens are processed by Hilbert Mamba. During training, ROI guidance is derived from the ground-truth segmentation mask and used to strengthen foreground-relevant tokens. During inference, no ROI mask or external proposal is used.

3.2 CONTENT-ADAPTIVE TOKENIZER

Uniform tokenization assigns the same representation budget to homogeneous background, organ interiors, lesion boundaries, and small abnormal structures, although these regions have very different segmentation difficulty. This is inefficient for 3D medical volumes, where clinically informative anatomy is often sparse. AAMamba therefore constructs content-adaptive region tokens before sequence modeling, rather than first forming a uniform grid and then pruning or merging tokens.

Given a stage feature map $x_n \in \mathbb{R}^{B\times C_n\times D_n\times H_n\times W_n}$, we compute a channel-averaged structural map $\bar{x}_n = \frac{1}{C_n}\sum_{c=1}^{C_n} x_{n,c}$ and estimate anatomical complexity using three cues: intensity deviation $d_{\text{int}} = |\bar{x}_n - \mu(\bar{x}_n)|$, boundary strength $d_{\text{grad}} = \|\nabla\bar{x}_n\|_2$, and curvature response $d_{\text{lap}} = |\Delta\bar{x}_n|$. With mean normalization $\text{nrm}(u) = u/(\text{mean}(u)+\epsilon)$, the structural density field is defined as

$$d_n = \text{ReLU}(\text{nrm}(d_{\text{int}}) + \text{nrm}(d_{\text{grad}}) + 0.5\,\text{nrm}(d_{\text{lap}}))\,. \tag{1}$$

High-density regions correspond to abnormal intensity variation, strong boundaries, or local structural irregularity, and therefore receive denser representation.

Figure 3 illustrates the tokenizer on a representative slice. The input image in Fig. 3(a) is converted into the density map in Fig. 3(b), where brighter responses indicate anatomically complex regions. The density field drives the relocation tendency shown in Fig. 3(c), producing the content-adaptive grid in Fig. 3(d). Let Q_n be the normalized 3D reference grid. We form the adapted grid as $Q'_n = Q_n + m_n\Delta_n$, where Δ_n is a displacement field and m_n is a density-dependent gate that suppresses unnecessary deformation in homogeneous areas. The adapted feature map is obtained by $\tilde{x}_n =$

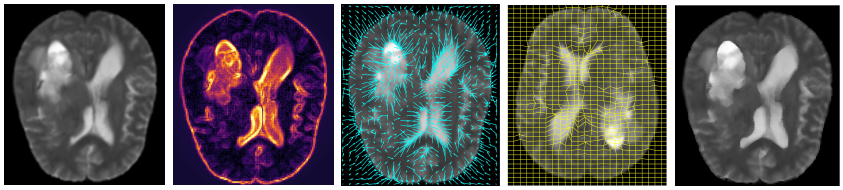

Figure 3: Content-adaptive tokenization on a representative slice. (a) Input slice. (b) Structural density field estimated from intensity, boundary, and curvature cues. (c) Density-guided relocation field. (d) Content-adaptive sampling grid, which allocates denser sampling around lesions and strong boundaries. (e) Regridded view before strided projection. The tokenizer produces adaptive tokens for sequence modeling while preserving a grid-aligned stream for coordinate-stable reconstruction.

$\mathrm{GridSample}(x_n, Q'_n)$, as visualized in Fig. 3(e). This regridding preserves full-volume coverage while increasing sampling emphasis around lesions, boundaries, and irregular structures.

The tokenizer outputs two aligned streams: $P_n = \mathrm{LN}(\psi_{\mathrm{flat}}(\mathrm{Conv}_s(\tilde{x}_n)))$ and $G_n = \mathrm{LN}(\psi_{\mathrm{flat}}(\mathrm{Conv}_s(x_n)))$, where Conv_s is a strided 3D projection and ψ_{flat} flattens the projected features into tokens. The adaptive stream P_n is used for anatomy-aware sequence modeling, while the grid-aligned stream G_n preserves an undistorted spatial reference for dense reconstruction. This dual-stream design allows AAMamba to allocate more representation to informative anatomy without losing coordinate consistency for voxel-level prediction.

3.3 ROI-Guided Bidirectional Hilbert Mamba

Mamba processes tokens as one-dimensional sequences; therefore, the scan order used to flatten 3D tokens directly affects spatial continuity. A fixed raster or Cartesian scan may separate neighboring 3D tokens and create artificial sequence neighbors across scan boundaries. Recent visual Mamba studies, such as DefMamba, show that predefined scan orders can weaken spatial-structure modeling and that adaptive scanning can prioritize informative regions Liu et al. (2025). Base on this insight, AAMamba propose ROI-Guided Bidirectional Hilbert Mamba segmentation. Rather than learning a general deformable scan path.

Given adaptive tokens $P_n \in \mathbb{R}^{B\times L_n\times C_n}$ at encoder stage n, we assign each token a 3D coordinate and sort the tokens using a Hilbert space-filling curve. Let π_n^+ denote the forward Hilbert order and $\pi_n^- = \mathrm{Reverse}(\pi_n^+)$ denote the reverse order. The serialized token sequences are

$$P_n^+ = P_n[\pi_n^+], \qquad P_n^- = P_n[\pi_n^-].$$

Hilbert ordering is used because it better preserves local 3D neighborhoods during flattening than simple raster or axis-wise traversal. The reverse scan is used to provide complementary context propagation, reducing the directional bias of a single sequence order.

The ROI guidance is derived from the training segmentation mask, not from an extra annotation. Given the ground-truth label map Y, we define the foreground ROI as $R = \mathbf{1}_{\{Y>0\}}$, i.e., the union of all foreground classes. For each encoder stage, R is downsampled to the corresponding token resolution to obtain R_n. As illustrated in Fig. 4, the scan is guided to emphasize tokens inside the lesion or organ ROI first, and then continues to the remaining non-ROI anatomy and background. Thus, the ROI provides training-time anatomical supervision while the full volume remains part of the sequence.

During training, the ROI map is reordered using the same forward and reverse Hilbert indices and applied as token-importance guidance:

$$\bar{P}_n^+ = P_n^+ \odot (1 + \alpha R_n[\pi_n^+]), \qquad \bar{P}_n^- = P_n^- \odot (1 + \alpha R_n[\pi_n^-]), \tag{2}$$

where α controls the guidance strength and $\odot$ denotes channel-wise broadcasting. This operation does not crop the volume and does not discard background tokens. Instead, it keeps the complete sequence while increasing the influence of lesion- or organ-relevant tokens during training.

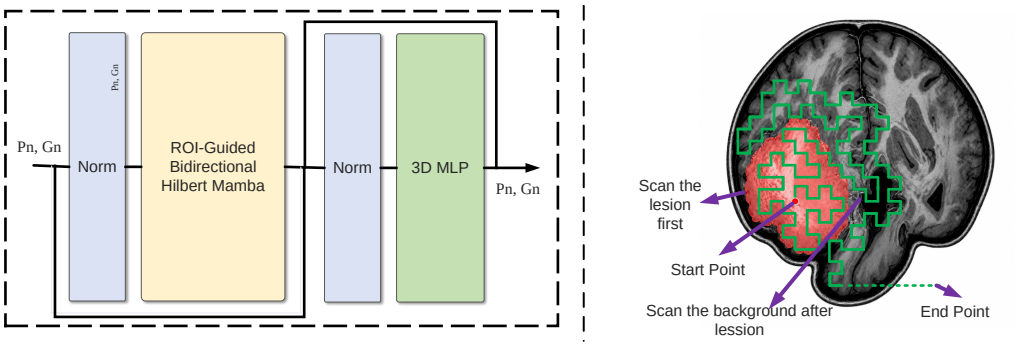

Figure 4: Detailed **ROI-guided Hilbert Mamba encoder block**. The **left** panel shows the encoder block used in the overall architecture: (P_n, G_n) are processed by normalization, ROI-guided Hilbert Mamba, normalization, and a 3D MLP. The **right** panel shows ROI-guided bidirectional Hilbert scanning, where the training-time foreground ROI guides the Hilbert sequence to traverse lesion- or organ-relevant tokens first and then continue through the surrounding non-ROI anatomy and background. This ROI-first ordering emphasizes clinically relevant regions while preserving full-volume context. No ROI input is required during inference.

The guided forward and reverse sequences are processed by Mamba blocks, restored to the original token layout with inverse permutations, and fused by residual integration:

$$Q_n^{\pm} = \mathrm{InvPerm}\Big(\mathrm{Mamba}\big(\mathrm{LN}(\bar{P}_n^{\pm})\big), \pi_n^{\pm}\Big). \tag{3}$$

$$\hat{P}_n = P_n + \mathrm{MLP}\Big(\mathrm{LN}\big[Q_n^{+}, Q_n^{-}\big]\Big). \tag{4}$$

The resulting feature $\hat{P}_n$ combines locality-preserving Hilbert serialization, bidirectional state-space propagation, and ROI-aware training guidance.

At inference, the ground-truth mask is unavailable and no ROI input is used. We set $R_n = 0$, so $\bar{P}_n^{+} = P_n^{+}$ and $\bar{P}_n^{-} = P_n^{-}$. Therefore, AAMamba receives only the input image volume at test time and predicts the segmentation mask directly.

3.4 Coordinate-Aware Decoder

Content-adaptive tokenization improves anatomical focus but changes the sampling geometry. Decoding only from adaptive features may therefore introduce spatial drift. The coordinate-aware decoder prevents this by fusing adaptive semantic features with grid-aligned spatial references.

For each stage, the processed adaptive tokens $\hat{P}_n$ and grid-aligned tokens G_n are reshaped into volumetric maps F_n^P and F_n^G. The two streams are fused as

$$F_n = \mathrm{Conv}_{1\times1\times1}\left([F_n^P, F_n^G]\right). \tag{5}$$

All stage features are upsampled to a common resolution and concatenated:

$$F_{\mathrm{cat}} = [\mathrm{Up}(F_1), \mathrm{Up}(F_2), \ldots, \mathrm{Up}(F_N)]. \tag{6}$$

An MLP head produces dense logits $S \in \mathbb{R}^{B\times K\times D\times H\times W}$:

$$S = \mathrm{MLP}_{\mathrm{head}}(\mathrm{LN}(F_{\mathrm{cat}})). \tag{7}$$

By using both adaptive features and the grid-aligned stream, the decoder preserves anatomical adaptivity while maintaining voxel-level coordinate consistency.

4 Experimental Results

4.1 Experimental Setup

We evaluate AAMamba on three standard 3D medical image segmentation benchmarks: BraTS for brain tumor segmentation, ACDC for cardiac MRI segmentation, and Synapse for multi-organ

Table 1: ACDC cardiac segmentation results. Dice similarity coefficient (DSC, %) is reported for right ventricle (RV), myocardium (Myo), and left ventricle (LV). Best results are bolded and second-best results are underlined.

Method	Params (M)	DSC	RV	Myo	LV
AAMamba	5.7	<u>91.85</u>	<u>90.00</u>	<u>89.92</u>	95.63
SegFormer3D Perera et al. (2024)	4.5	90.96	88.50	88.86	95.53
LeViT-UNet-384 Xu et al. (2023)	52.17	90.32	89.55	87.64	93.76
SwinUNet	–	90.00	88.55	85.62	<u>95.83</u>
TransUNet Chen et al. (2021)	96.07	89.71	88.86	84.54	95.73
UNETR Hatamizadeh et al. (2022b)	92.49	88.61	85.29	86.52	94.02
R50-ViT-CUP Chen et al. (2021)	86.00	87.57	86.07	81.88	94.75
R50 ViT Chen et al. (2021)	–	87.57	86.07	81.88	94.75
R50 U-Net Chen et al. (2021)	–	87.55	87.10	80.63	94.92
R50 Att-UNet Chen et al. (2021)	–	86.75	87.58	79.20	93.47
ViT-CUP Chen et al. (2021)	86.00	81.45	81.46	70.71	92.18

Table 2: BraTS tumor segmentation results. Dice score (%) is reported for Whole Tumor (WT), Enhancing Tumor (ET), and Tumor Core (TC). Best results are bolded and second-best results are underlined.

Method	Params (M)	Mean	WT	ET	TC
AAMamba	5.7	<u>84.2</u>	<u>90.9</u>	<u>77.1</u>	<u>84.7</u>
SegFormer3D Perera et al. (2024)	4.5	82.1	89.9	74.2	82.2
SegMamba Xing et al. (2025)	67.4	81.2	90.3	71.9	81.4
UNETR Hatamizadeh et al. (2022b)	92.5	71.1	78.9	58.5	76.1
TransBTS Wang et al. (2021)	–	69.6	77.9	57.4	73.5
CoTr Xie et al. (2021)	41.9	68.3	74.6	55.7	74.8
TransUNet Chen et al. (2021)	96.1	64.4	70.6	54.2	68.4
SETR-MLP Zhang & Zhang (2022)	310.5	63.9	69.8	55.4	66.5
SETR-PUP Zhang & Zhang (2022)	318.3	63.8	69.6	54.9	67.0
SETR-NUP Zhang & Zhang (2022)	305.7	63.7	69.7	54.4	66.9

abdominal CT segmentation. Following common evaluation protocols, we report Dice score (%) for each anatomical target and use mean Dice or DSC as the primary accuracy metric. We report model efficiency using the number of trainable parameters, and conduct all training and inference experiments on an NVIDIA GeForce RTX 5090 GPU.

As shown in Fig. 1, AAMamba achieves a favorable accuracy–efficiency trade-off across BraTS, ACDC, and Synapse. With only 5.7M parameters, it remains compact while improving segmentation of small, thin, and morphologically irregular structures.

4.2 Quantitative Results

BraTS. Table 2 reports tumor segmentation performance on WT, ET, and TC. AAMamba achieves 84.2% mean Dice with 5.7M parameters. Compared with the compact SegFormer3D baseline, AAMamba improves mean Dice by $+2.1$ points. The gain is particularly clear for ET, where Dice increases from 74.2% to 77.1%. Since enhancing tumor is typically small and irregular, this improvement supports the role of content-adaptive token allocation and locality-preserving sequence construction in preserving fine lesion evidence.

ACDC. Table 1 shows cardiac segmentation results on RV, myocardium, and LV. AAMamba obtains 91.85% DSC with only 5.7M parameters. Compared with SegFormer3D, AAMamba improves the overall DSC by $+0.89$ points. The myocardium result is especially informative: AAMamba improves Myo Dice from 88.86% to 89.92%, achieving the best myocardium performance among the listed methods. Since myocardium is a thin, low-contrast structure, this result suggests that the proposed representation better preserves boundary-sensitive anatomical detail.

Table 3: Synapse multi-organ segmentation results. Dice score (%) is reported for aorta (AOR), liver (LIV), left kidney (LKID), right kidney (RKID), gallbladder (GAL), pancreas (PAN), spleen (SPL), and stomach (STO). Avg denotes the average Dice score. Best results are bolded and second-best results are underlined.

Method	Params	Avg	AOR	LIV	LKID	RKID	GAL	PAN	SPL	STO
AAMamba	5.7	<u>84.26</u>	<u>89.78</u>	<u>94.61</u>	<u>84.23</u>	<u>83.94</u>	68.01	<u>80.96</u>	88.37	<u>84.15</u>
SegFormer3D Perera et al. (2024)	4.5	82.15	90.43	95.68	86.53	86.13	55.26	73.06	89.02	81.12
MISSFormer Huang et al. (2021)	–	81.96	86.99	94.41	85.21	82.00	<u>68.65</u>	65.67	<u>91.92</u>	80.81
UNETR Hatamizadeh et al. (2022b)	92.49	79.56	89.99	94.46	85.66	84.80	60.56	59.25	87.81	73.99
SwinUNet Hatamizadeh et al. (2022a)	–	79.13	85.47	94.29	83.28	79.61	66.53	56.58	90.66	76.60
LeViT-UNet-384 Xu et al. (2023)	52.17	78.53	87.33	93.11	84.61	80.25	62.23	59.07	88.86	72.76
TransClaw U-Net Chang et al. (2021)	–	78.09	85.87	94.28	84.83	79.36	61.38	57.65	87.74	73.55
TransUNet Chen et al. (2021)	96.07	77.48	87.23	94.08	81.87	77.02	63.16	55.86	85.08	75.62
R50-ViT-CUP Chen et al. (2021)	86.00	71.29	73.73	91.51	75.80	72.20	55.13	45.99	81.99	73.95
ViT-CUP Chen et al. (2021)	86.00	67.86	70.19	91.32	74.70	67.40	45.10	42.00	81.75	70.44

Table 4: Component ablation. CAT denotes content-adaptive tokenization, GR denotes the grid-aligned stream, and Hilbert denotes locality-preserving token ordering. Dice scores (%) are reported.

Variant	BraTS		ACDC		Synapse	
	Mean	ET	DSC	Myo	Avg	GAL/PAN
w/o CAT (uniform)	82.9	74.8	91.05	88.95	82.90	60.4/74.1
w/o GR	83.7	76.0	91.55	89.45	83.75	66.3/79.1
w/o Hilbert (raster)	84.0	76.5	91.72	89.70	84.12	67.6/80.3
Full	**84.2**	**77.1**	**91.85**	**89.92**	**84.26**	**68.01/80.96**

Synapse. Table 3 reports multi-organ segmentation results on Synapse. AAMamba achieves 84.26% average Dice with 5.7M parameters. Compared with SegFormer3D, AAMamba improves the average Dice by $+2.11$ points. The gains are strongest on small and irregular organs: gallbladder improves from 55.26% to 68.01%, and pancreas improves from 73.06% to 80.96%. These improvements are consistent with the motivation of AAMamba: adaptive token allocation benefits sparse and morphologically complex structures, while coordinate-aware reconstruction helps preserve dense voxel-level spatial consistency.

4.3 Ablation Study

We isolate the contribution of each component while keeping the training protocol, data splits, and backbone capacity fixed. All variants use the same encoder–decoder topology and 3D state-space backbone, and differ only by removing one component at a time: content-adaptive tokenization (CAT), the grid-aligned reconstruction stream (GR), and Hilbert ordering for sequence construction. We report representative metrics that are sensitive to small or thin structures: ET on BraTS, Myo on ACDC, and GAL/PAN on Synapse.

Table 4 shows that each component contributes to the final performance. Removing CAT produces the largest degradation, reducing BraTS mean Dice from 84.2% to 82.9%, ACDC DSC from 91.85% to 91.05%, and Synapse average Dice from 84.26% to 82.90%. The drop is most pronounced on ET, GAL, and PAN, supporting the role of adaptive token construction in allocating more representation to small lesions, boundaries, and irregular organs. Removing GR also consistently degrades performance, indicating that decoding only from adaptive features can introduce spatial drift during voxel-level reconstruction. Replacing Hilbert ordering with raster ordering yields a smaller but consistent drop, suggesting that locality-preserving serialization improves state-space propagation by reducing artificial discontinuities introduced by naive flattening.

Sensitivity to adaptive resolution. Table 5 evaluates the number of content-adaptive control points P. Increasing P from 64 to 128 improves performance across all datasets, while increasing P further to 256 provides no additional benefit. This suggests that moderate adaptive resolution is sufficient to capture stable anatomical structure, whereas excessively dense control points may become more sensitive to local intensity noise.

Table 5: Sensitivity to the number of content-adaptive control points P. The best setting is bolded.

Control points P	BraTS Mean	ACDC DSC	Synapse Avg
64	84.0	91.74	84.10
128	**84.2**	**91.85**	**84.26**
256	84.1	91.83	84.22

5 Conclusion

This work presented **AAMamba**, an anatomy-aware Mamba framework for efficient adaptive 3D medical image segmentation. AAMamba addresses the volume-to-sequence construction problem in 3D state-space segmentation by jointly designing content-adaptive token allocation, ROI-guided Hilbert sequence modeling, and coordinate-aware dense reconstruction. Instead of relying on content-agnostic volumetric tokenization, AAMamba constructs adaptive region tokens according to anatomical complexity, allowing the encoder to allocate greater representation to lesions, boundaries, and irregular structures while maintaining full-volume coverage. ROI-guided bidirectional Hilbert scanning further improves Mamba-based sequence modeling by prioritizing lesion- or organ-relevant tokens during training, while requiring no ROI input during inference. To recover voxel-level predictions without spatial drift, the coordinate-aware decoder fuses anatomy-adaptive semantic features with grid-aligned spatial references. Experiments on BraTS, ACDC, and Synapse show that AAMamba achieves a strong compact-model accuracy–efficiency trade-off, obtaining 84.2% mean Dice on BraTS, 91.85% DSC on ACDC, and 84.26% average Dice on Synapse with only 5.7M parameters. These results support anatomy-aware sequence construction as an effective representation strategy for compact 3D medical state-space segmentation, especially for small or morphologically complex anatomical structures.

6 Limitations

The current study targets efficient 3D medical image segmentation, and therefore focuses on compact and moderate-size baselines with fewer than 100M parameters. Very high-capacity models such as nnFormer variants, and other models with more than 100M parameters are not included in the main experimental comparison because they operate in a different capacity regime and are primarily optimized for absolute accuracy rather than compact deployment. Consequently, the present experiments should be interpreted as evidence for the accuracy–efficiency advantage of AAMamba in the lightweight setting, rather than as a claim of absolute state-of-the-art performance against all large-scale 3D segmentation systems.

References

Yao Chang, Hu Menghan, Zhai Guangtao, and Zhang Xiao-Ping. Transclaw u-net: Claw u-net with transformers for medical image segmentation. *arXiv preprint arXiv:2107.05188*, 2021.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2016.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

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. *arXiv preprint arXiv:2010.11929*, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *arXiv preprint arXiv:2201.01266*, 2022a.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2022b.

Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Leiye Liu, Miao Zhang, Jihao Yin, Tingwei Liu, Wei Ji, Yongri Piao, and Huchuan Lu. Defmamba: Deformable visual state space model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8838–8847, 2025.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

Dong Lu, Long Xu, Jie Zhou, Kai Gao, Zhiwei Gong, and Di Zhang. 3d-umamba: 3d u-net with state space model for semantic segmentation. *International Journal of Applied Earth Observation and Geoinformation*, 136:104401, 2025.

Jian Ma, Feng Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Shehan Perera, Pouyan Navard, and Alper Yilmaz. Segformer3d: An efficient transformer for 3d medical image segmentation. In *CVPR Workshops (DEF-AI-MIA)*, 2024.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Yongming Rao, Wenliang Zhao, Zhiyuan Zhu, Jiwen Lu, Jie Zhou, and Chih-Yao Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. Springer, 2015.

Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.

Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Sudhakar Sah, Ravish Kumar, Honnesh Rohmetra, and Ehsan Saboori. Token pruning using a lightweight background aware vision transformer. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*, 2024. URL https://openreview.net/forum?id=6ZaSMpmNVF.

Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International conference on medical image computing and computer-assisted intervention*, pp. 109–119. Springer, 2021.

Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 171–180. Springer, 2021.

Zhaohu Xing, Tian Ye, Yijun Yang, Du Cai, Baowen Gai, Xiao-Jian Wu, Feng Gao, and Lei Zhu. Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

Guoping Xu, Xuan Zhang, Xinwei He, and Xinglong Wu. Levit-unet: Make faster encoders with transformer for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 42–53. Springer, 2023.

Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24972–24982, 2025.

Rui Zhang and Shihua Zhang. Rethinking influence functions of neural networks in the over-parameterized regime. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9082–9090, 2022.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

Liang Zhu et al. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

Appendix

This appendix provides additional qualitative evidence and architectural motivation for AAMamba. Section A compares AAMamba with representative baselines on BraTS, ACDC, and Synapse, highlighting challenging cases involving small lesions, thin boundaries, and irregular organs. Section **??** explains why ROI-guided Hilbert scanning is used for 3D state-space segmentation, including the rationale for lesion-first traversal, the limitation of raster scanning, and the locality-preserving role of Hilbert ordering. These analyses support the main claim that anatomy-aware volume-to-sequence construction improves compact 3D medical segmentation.

A Qualitative Results

We provide qualitative comparisons with SegFormer3D, nnFormer, and UNETR on three representative datasets. These examples are selected to visualize the types of structures where AAMamba provides the largest empirical gains in the main paper. In each figure, the first column shows the ground-truth annotation, the middle columns show competing methods, and the final column shows AAMamba. Highlighted boxes indicate regions where competing methods produce missing structures, fragmented masks, false positives, or boundary inconsistencies.

A.1 BraTS

BraTS contains tumor subregions with highly irregular shapes and large variation in size. In particular, enhancing tumor regions are often small, discontinuous, and difficult to distinguish from surrounding abnormal tissue. Figure 5 shows that competing methods may miss small tumor components or produce fragmented tumor-core predictions in boundary-sensitive regions. AAMamba produces more spatially coherent tumor masks and better preserves small enhancing structures. This behavior is consistent with the quantitative improvement on ET reported in the main paper.

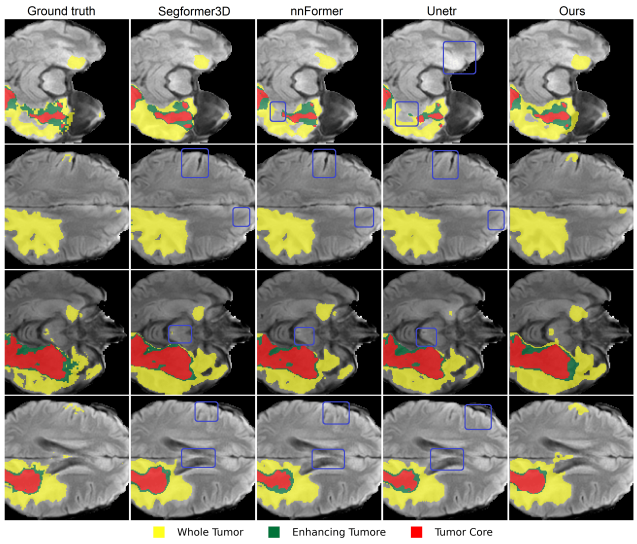

Figure 5: Qualitative comparison on BraTS. Columns show ground truth, SegFormer3D, nnFormer, UNETR, and AAMamba. Colors denote tumor subregions: whole tumor, enhancing tumor, and tumor core. Blue boxes highlight regions where competing methods miss small enhancing components, produce fragmented tumor predictions, or introduce boundary inconsistencies. AAMamba better preserves small and irregular tumor structures.

The BraTS examples suggest that anatomy-aware sequence construction helps the model focus on sparse tumor evidence without losing global context. Content-adaptive tokenization increases representation around abnormal intensity and boundary regions, while ROI-guided Hilbert scanning encourages foreground-relevant tokens to influence state propagation more strongly during training.

This is particularly useful for enhancing tumor, which occupies a small portion of the volume but strongly affects clinical segmentation quality.

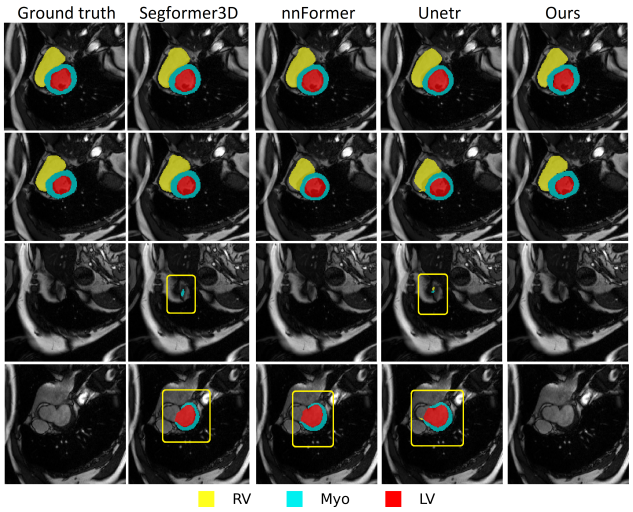

Figure 6: Qualitative comparison on ACDC cardiac MRI. Columns show ground truth, SegFormer3D, nnFormer, UNETR, and AAMamba. Colors denote right ventricle (RV), myocardium (Myo), and left ventricle (LV). Yellow boxes highlight boundary-sensitive regions where competing methods produce false positives, missing components, or unstable myocardium contours. AAMamba produces more coherent masks, especially around thin myocardial boundaries.

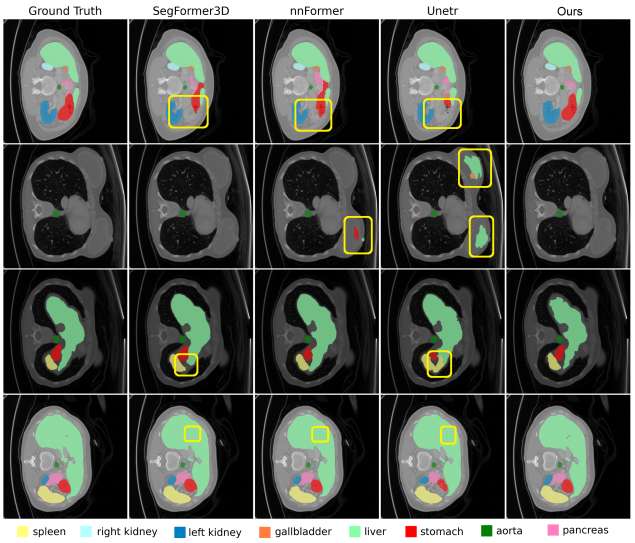

Figure 7: Qualitative comparison on Synapse multi-organ segmentation. Columns show ground truth, SegFormer3D, nnFormer, UNETR, and AAMamba. Colors denote abdominal organs including spleen, kidneys, gallbladder, liver, stomach, aorta, and pancreas. Yellow boxes highlight regions where competing methods introduce fragmented predictions, false positives, or missing small organs. AAMamba yields more stable organ boundaries and better preserves small or irregular structures such as gallbladder and pancreas.

A.2 ACDC

ACDC cardiac MRI segmentation requires accurate delineation of the right ventricle, myocardium, and left ventricle. Among these structures, the myocardium is especially challenging because it

forms a thin ring-like structure with low-contrast boundaries. Figure 6 shows representative cases where competing methods produce small false positives, missing regions, or less stable myocardial contours. AAMamba yields more coherent cardiac masks and better preserves the relative structure of RV, Myo, and LV.

The ACDC results illustrate the importance of coordinate-aware reconstruction. Since adaptive tokenization changes the sampling geometry, decoding only from adaptive features may lead to local spatial drift. AAMamba mitigates this issue by retaining a grid-aligned stream and fusing it with anatomy-adaptive features in the decoder. This helps preserve the spatial consistency needed for thin cardiac structures.

A.3 SYNAPSE

Synapse multi-organ segmentation contains organs with different sizes, shapes, and contrast characteristics. Large organs such as the liver are relatively easier to segment, while small or irregular organs such as gallbladder and pancreas are more sensitive to boundary errors and missing predictions. Figure 7 shows that competing methods sometimes introduce fragmented masks, false positives, or missing small-organ regions. AAMamba provides more stable segmentation for small and irregular structures while maintaining coherent masks for larger organs.