SAMba-UNet: Synergizing SAM2 and Mamba in UNet with Heterogeneous Aggregation for Cardiac MRI Segmentation

Guohao Huo
School of Information and Software Engineering
University of Electronic Science and Technology of China
gh.huo513@gmail.com

Ruiting Dai
School of Information and Software Engineering
University of Electronic Science and Technology of China
rtdai@uestc.edu.cn

Hao Tang *
Peking University
hao.tang@vision.ee.ethz.ch

Abstract

To address the challenge of complex pathological feature extraction in automated cardiac MRI segmentation, this study proposes an innovative dual-encoder architecture named SAMba-UNet. The framework achieves cross-modal feature collaborative learning by integrating the vision foundation model SAM2, the state-space model Mamba, and the classical UNet. To mitigate domain discrepancies between medical and natural images, a Dynamic Feature Fusion Refiner is designed, which enhances small lesion feature extraction through multi-scale pooling and a dual-path calibration mechanism across channel and spatial dimensions. Furthermore, a Heterogeneous Omni-Attention Convergence Module (HOACM) is introduced, combining global contextual attention with branch-selective emphasis mechanisms to effectively fuse SAM2's local positional semantics and Mamba's long-range dependency modeling capabilities. Experiments on the ACDC cardiac MRI dataset demonstrate that the proposed model achieves a Dice coefficient of 0.9103 and an HD95 boundary error of 1.0859 mm, significantly outperforming existing methods, particularly in boundary localization for complex pathological structures such as right ventricular anomalies. This work provides an efficient and reliable solution for automated cardiac disease diagnosis, and the code will be open-sourced.

1 Introduction

Cardiovascular diseases remain one of the leading causes of global mortality, with early diagnosis predominantly relying on imaging modalities such as cardiac magnetic resonance imaging (MRI). However, conventional MRI analysis requires manual annotation of cardiac structures (e.g., ventricles, myocardium) by specialized clinicians, a process that is time-consuming and prone to subjective variability. This limitation becomes particularly evident in detecting complex pathologies like ischemic heart failure, hypertrophic cardiomyopathy, and right ventricular anomalies, where human interpretation struggles to achieve high sensitivity and consistency. Driven by transformative advances in artificial intelligence (AI), the medical community is now developing automated algorithmic

[1] *Corresponding author.

Preprint. Under review.

systems to enhance diagnostic efficiency Sanyaolu [2025], minimize human error, and deliver data-driven decision support for early screening, personalized treatment optimization, and prognostic evaluation of cardiac disorders.

In recent years, deep learning architectures have achieved remarkable breakthroughs in medical image segmentation, with encoder-decoder based convolutional neural networks (CNNs) demonstrating exceptional performance Ronneberger et al. [2015], Zhou et al. [2019]. As a milestone architecture in this field, UNet has established its core position in medical image segmentation tasks through its unique symmetric encoder-decoder design and cross-level skip connection mechanism. To enhance feature representation capabilities, subsequent studies have developed various innovative auxiliary modules Huang et al. [2017], He et al. [2016], Woo et al. [2018], Howard et al. [2017], Yu and Koltun [2015], Zhou et al. [2020]. These technological advancements have empowered the UNet architecture to demonstrate outstanding clinical value in segmenting medical images across multiple modalities, including CT, MRI, and ultrasound imaging.

Transformer demonstrates significant advantages in modeling long-range dependencies and capturing global context through its attention mechanism Chen et al. [2021], Zhang et al. [2021], Valanarasu et al. [2021], Hatamizadeh et al. [2021]. Driven by advanced network architectures Dosovitskiy et al. [2020] and large-scale datasets Kirillov et al. [2023], recent segmentation trends have shifted from task-specific expert models toward general-purpose foundation models that can perform segmentation without extensive task-specific development Moor et al. [2023], He et al. [2024], Khan et al. [2025]. SAM Kirillov et al. [2023] and SAM2 Ravi et al. [2024], as newly developed visual foundation models, have shown impressive zero-shot performance across various natural image tasks. However, the substantial domain gap between natural images and MRI scans prevents direct deployment of SAM on medical imaging Huang et al. [2024], He et al. [2023], Roy et al. [2023]. To address the challenges of blurred or missing small lesions and fine structures when applying SAM2 to MRI segmentation due to its training on natural images, we propose a Dynamic Feature Fusion Refiner in this study.

Although the Transformer architecture demonstrates remarkable advantages in modeling long-range dependencies, its inherent quadratic computational complexity leads to excessive resource consumption in medical image segmentation tasks Liu et al. [2024a], Huang et al. [2022]. In contrast, the Mamba architecture emerges as a promising solution for medical imaging due to its linear computational complexity and powerful capability in capturing long-range dependencies Xing et al. [2024], Yang et al. [2024]. In the Hiera architecture employed by SAM2, the window attention mechanism and window-based absolute positional encoding may cause loss of pixel-level spatial positional semantic information, thereby compromising the extraction of segmentation boundary features. To enhance long-range dependency learning while reducing computational resource consumption, we introduce the VMamba architecture to capture global semantic features that complement SAM2's hierarchical features. Addressing the challenge of adaptive fusion between SAM2 and Mamba encoder features, we innovatively design a Heterogeneous Omni-Attention Convergence Module (HOACM) to effectively integrate heterogeneous semantic features from both architectures. Based on our collection and analysis of existing data, it indicates that the model we have proposed, SAMba-UNet represents the first pioneering framework that successfully synergizes SAM2, Mamba, and UNet architectures.

In conclusion, our contributions are as follows: (1) Propose the first synergistic framework (SAMba-UNet) integrating the visual foundation model (SAM2), state space model (Mamba), and classical UNet, resolving the trade-off between global semantic modeling and local detail capture in medical image segmentation. (2) Design a multi-scale pooling and channel-spatial dual-path calibration module called Dynamic Feature Fusion Refiner to mitigate domain gaps between natural and medical images, enhancing segmentation robustness for small lesions. (3) introduced the Heterogeneous Omni-Attention Convergence Module (HOACM). Develop a cross-architecture attention fusion mechanism: OCA strengthens pixel-level positional semantics, while BSEA enables dynamic aggregation of global-local features (4) SAMba-UNet achieves a Dice score of 0.9103 and HD95 boundary error of 1.0859mm on the ACDC cardiac MRI dataset, establishing a new technical standard for clinical cardiac function quantification.

2 Related Work

2.1 Medical Image Segmentation with SAM

The introduction of the Segment Anything Model (SAM) Kirillov et al. [2023] represents a significant milestone in image segmentation. While SAM demonstrates remarkable zero-shot segmentation capabilities through user prompts (e.g., points, bounding boxes) without task-specific training, its direct application to medical imaging, particularly MRI analysis, faces challenges due to substantial domain discrepancies between natural images and medical imaging modalities. To address this, parameter-efficient fine-tuning (PEFT) Zu et al. [2024] strategies offer an effective solution by updating minimal parameters (typically <5% of total weights) while maintaining most parameters frozen. Building on this concept, Wu et al. [2025] proposes a Medical SAM Adapter (Med-SA) that injects medical domain knowledge through PEFT rather than direct SAM fine-tuning, achieving state-of-the-art performance in medical image segmentation with only 2% parameter updates. Further optimizing for MRI segmentation tasks, we introduce a Dynamic Feature Fusion Refiner module to address the limitations of SAM in capturing small lesions and subtle anatomical structures, which stems from its extensive training on natural image datasets.

2.2 Medical Image Segmentation with Mamba

State space sequence models (SSMs) such as Mamba Gu and Dao [2023] offer a novel approach for efficient global dependency modeling through their linear complexity (O(n)) in long sequence processing. Unlike self-attention mechanisms, SSMs achieve interaction between sequence elements and historical information by compressing hidden states, thereby avoiding quadratic computational overhead. The Mamba-Unet Wang et al. [2024] framework proposes a novel medical image segmentation model by integrating the U-Net Ronneberger et al. [2015] architecture with Mamba's capabilities. However, its insufficient local feature extraction capability limits effective capture of subtle lesion structures. Subsequent improvements like U-Mamba Ma et al. [2024] and SegMamba Xing et al. [2024] combine Mamba with CNNs for direct pixel-level long-range dependency modeling. Nevertheless, these Mamba-based models compromise the spatial continuity of local neighborhood pixels due to their 1D sequential processing, adversely affecting detail modeling. To address these limitations, we propose leveraging SAM2's Ravi et al. [2024] capability in capturing window-based absolute positional spatial semantic information to compensate for the detailed modeling deficiencies inherent in Mamba architectures.

2.3 Synergizing Segment Anything Model with Mamba Architecture

SAM-Mamba Dutta et al. [2025] proposes a Mamba-guided Segment Anything Model for efficient polyp segmentation, introducing a Mamba-Prior module as a bridge to connect SAM's general pre-trained representations with polyp-related subtle cues. LFSamba Liu et al. [2024b] develops a novel multi-focus light field image salient object detection model that reconstructs 3D scenes using single-focus images to capture spatial geometric information. While existing works integrate SAM with Mamba, this study innovatively combines SAM2 with Mamba. SAM2's MAE He et al. [2022] -pretrained hierarchical Hiera image encoder Ryali et al. [2023] completely removes relative position bias (RPB) during attention computation, adopting window-based absolute position encoding Yu et al. [2024], which may compromise pixel-level spatial positional semantics and adversely affect segmentation performance. To address this, we propose leveraging Mamba's significant advantage in capturing global semantic information with linear computational complexity, thereby synergistically enhancing SAM2's capacity to model both global semantics and position-sensitive local features.

3 Method

Samba-UNet is a U-shaped architecture featuring dual-stream encoders (SAM2 and VMamba) and a single VMamba decoder. In the SAM2 encoder branch, we employ a Dynamic Feature Fusion Refiner with an MLP-Adapter to adaptively refine multi-scale features from the frozen SAM2 Hiera-Large encoder. The VMamba encoder branch utilizes state space modeling to enhance global semantic feature extraction, complementing SAM2's local focus. A novel Heterogeneous Omni-Attention Convergence Module dynamically fuses cross-architecture features from both encoders through

attention-based enhancement, feeding the integrated representations to the VMamba decoder for final segmentation prediction. The model architecture diagram is shown in Figure 1.

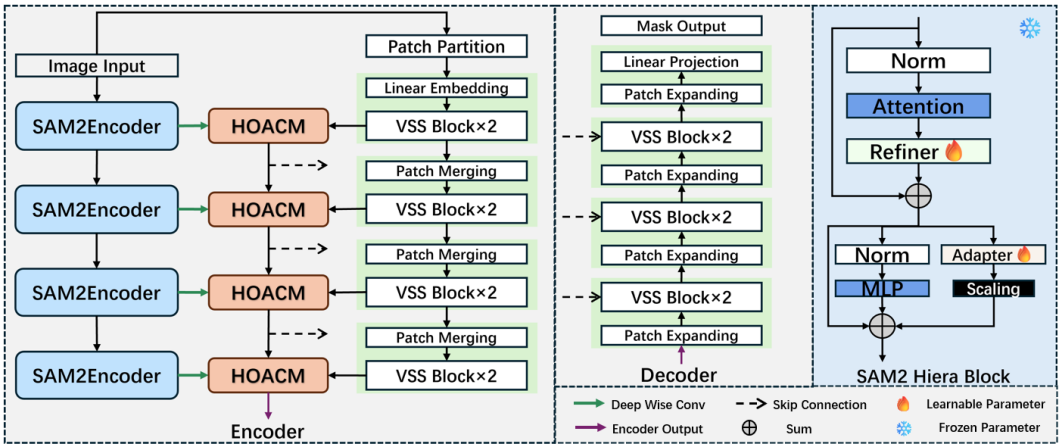

Figure 1: The architecture of SAMba-UNet.

3.1 Architecture Overview

Dual-stream encoders In the SAM2 encoder branch, we integrate design principles from FE-UNet and Medical SAM Adapter to establish a dual-Adapter architecture for Hiera Block fine-tuning. To bridge the domain gap between SAM2's natural image pre-training and MRI characteristics, we propose a Dynamic Feature Fusion Refiner that performs domain-adaptive refinement on attention outputs, complemented by parallel MLP-Adapter operations to jointly enhance nonlinear mapping capacity.

In the Mamba encoder branch, we leverage the same VSS Block configuration as Mamba-UNet to capture global semantic contexts, thereby effectively addressing the pixel-level spatial positional semantics loss in SAM2 caused by its windowing position encoding mechanism.

To achieve effective fusion of SAM2 and VMamba encoder features, we propose the Heterogeneous Omni-Attention Convergence Module (HOACM). This module integrates two core components: 1) Omniscient Contextual Attention (OCA) that enhances SAM2's pixel-level spatial-semantic relationship modeling through global contextual awareness, and 2) Bifurcated Selective Emphasis Attention (BSEA) that enables adaptive channel-spatial co-enhancement of VMamba features for cross-architecture dynamic aggregation.

Single VMamba decoder In the decoder design, we maintain architectural consistency with Mamba-UNet by implementing its progressive upsampling framework. This multi-stage feature aggregation mechanism enables hierarchical feature fusion through successive transpose convolutions, ultimately producing high-resolution segmentation masks via parameter-shared prediction heads.

3.2 Dynamic Feature Fusion Refiner

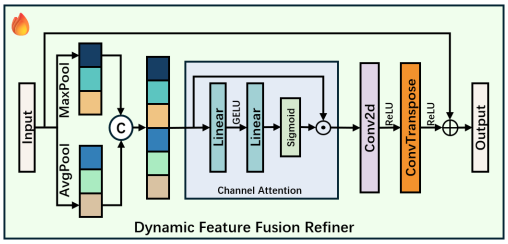

Figure 2: The architecture of Dynamic Feature Fusion Refiner.

Given the domain discrepancy between SAM2's natural image pre-training and MRI characteristics (including inherent blurring of details, attenuation of small lesion/subtle structure features due to physical imaging mechanisms, and non-rigid deformations caused by patient motion that compromise anatomical structural relationships), we propose the Dynamic Feature Fusion Refiner (The module architecture is illustrated in Figure 2) to enhance medical image adaptation capabilities.

The Dynamic Feature Fusion Refiner takes attention outputs as input features $X \in R^{B\times H\times W\times C}$. It employs dual adaptive pooling operations: adaptive max pooling preserves globally salient semantics while adaptive average pooling suppresses high-frequency noise interference. These pooled features are concatenated along the channel dimension after dimension permutation:

$$X_{\text{cat}} = \text{Concat}([\text{AdaptiveAvgPool2d}(\text{Permute}(X)), \text{AdaptiveMaxPool2d}(\text{Permute}(X))], \text{dim} = 1) \tag{1}$$

The concatenated feature $X_{\text{cat}} \in R^{B\times 2C}$ undergoes channel-wise dynamic calibration via a learnable attention gating mechanism. This bottleneck-structured operation models inter-channel dependencies to selectively amplify discriminative channels while suppressing noise-corrupted ones:

$$X_{\text{attn}} = X_{\text{cat}} \odot \sigma\left(W_2^{(c)} \cdot \text{ReLU}(W_1^{(c)} \cdot X_{\text{cat}})\right) \tag{2}$$

The parameters $W_1^{(c)} \in \mathbb{R}^{h\times 2C}$ and $W_2^{(c)} \in \mathbb{R}^{C\times h}$ in the formula represent learnable weight matrices, where $h = \lfloor C \cdot r \rfloor$ denotes the bottleneck dimension with compression ratio $r \in [0, 1]$, and σ serves as the channel gating activation. A cascaded convolution path is constructed to augment local receptive fields: the downsampling convolution compresses spatial dimensions, while the transposed convolution restores resolution, with ReLU non-linearities enhancing local feature extraction:

$$X_{\text{sp}} = \text{ReLU}\left(\text{DeConv2D}(\text{ReLU}\left(\text{Conv2D}(X_{\text{att}}, \mathbf{K}_d)\right), \mathbf{K}_u)\right) \tag{3}$$

The refined features are integrated with original inputs through residual connections to preserve multi-scale information, while Layer Normalization is applied to stabilize gradient propagation and ensure numerical stability during training:

$$X_{\text{out}} = \text{LayerNorm}(\text{Permute}(X_{\text{cat}} + X_{\text{sp}})) \tag{4}$$

3.3 Heterogeneous Omni-Attention Convergence Module

To address cross-level semantic discrepancies and enable effective fusion between SAM2 and VMamba encoder outputs, we design the Heterogeneous Omni-Attention Convergence Module (HOACM) (The module architecture is illustrated in Figure 3). This module comprises three key components: 1) Omniscient Contextual Attention (OCA) that enhances SAM2's global semantic representation through multi-scale contextual awareness, 2) Bifurcated Selective Emphasis Attention (BSEA) employing channel-spatial dual pathways to dynamically amplify discriminative VMamba features, and 3) a cross-attention fusion mechanism that hierarchically integrates local details, global contexts, and historical fused semantics via staged feature interactions.

Bifurcated Selective Emphasis Attention(BSEA) The Bifurcated Selective Emphasis Attention (BSEA) employs a dual-path pooling-attention architecture: 1) A spatial average pooling pathway preserves VMamba's global context modeling strength by capturing cross-region long-range dependencies, while 2) a local saliency pooling pathway applies spatial constraints to prevent attention over-focusing, thereby alleviating local-global semantic misalignment.

$$X_{avg} = \text{AvgPool}(X_{mamba}) \tag{5}$$

Given the Mamba encoder's layer-wise output features $X_{mamba} \in R^{B\times C\times H\times W}$, we implement a parallel spatial max-pooling pathway to preserve VMamba's inherent strength in capturing locally salient patterns. This design enhances sensitivity to critical segmentation boundaries by emphasizing regional response extremum features:

$$X_{max} = \text{MaxPool}(X_{mamba}) \tag{6}$$

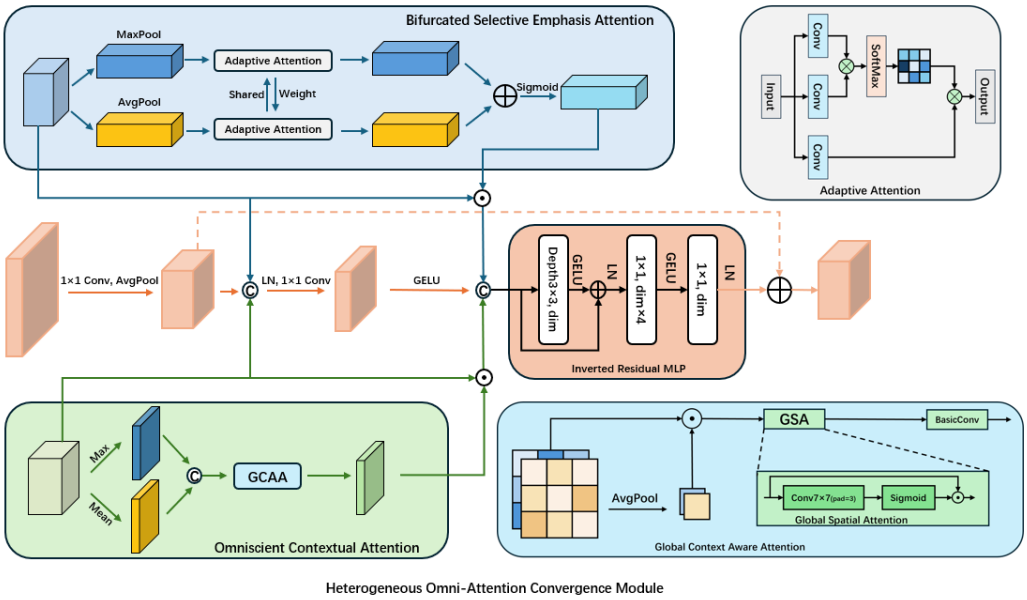

Figure 3: The architecture of HOACM.

The pooled features $X_{avg/max} \in R^{B\times C\times H'\times W'}$ are processed by a shared-weight adaptive self-attention mechanism to achieve inter-path feature space alignment and dynamic contribution balancing. Specifically, three independent 1×1 convolutional layers generate the Query, Key, and Value triplets:

$$Q = \text{Flatten}(\text{Conv2D}_{1\times1}(X_{avg/max}; W_q)) \tag{7}$$

$$K = \text{Flatten}(\text{Conv2D}_{1\times1}(X_{avg/max}; W_k)) \tag{8}$$

$$V = \text{Flatten}(\text{Conv2D}_{1\times1}(X_{avg/max}; W_v)) \tag{9}$$

The convolution kernels W_q, W_k, W_v in the formula are learnable parameters that generate corresponding $Q, K, V \in R^{B\times C\times N}$ (where $N = H \times W$ denotes flattened spatial dimensions). The similarity matrix between Query and Key is computed, followed by Softmax normalization to obtain spatial relationship weights:

$$M = \text{Softmax}\left(Q \times K^{\top},\ \text{dim} = -1\right) \tag{10}$$

The attention weight matrix M dynamically aggregates contextual information through matrix multiplication with the Value features V :

$$X'_{avg/max} = M \times V \tag{11}$$

The dual-path features processed by the shared-weight adaptive self-attention mechanism are fused through weighted summation to produce spatial attention maps. These maps enhance the original Mamba features via element-wise multiplication:

$$M = \text{Softmax}\left(X'_{avg} + X'_{max}\right) \times X_{mamba} \tag{12}$$

Omniscient Contextual Attention To address the SAM2 encoder's limitations in modeling global pixel-level positional semantics, we propose the Omniscient Contextual Attention (OCA) mechanism. This module enhances feature representation through a global contextual awareness attention architecture that establishes long-range cross-region dependencies and enables multi-scale semantic integration:

The mechanism initially performs dual-channel compression operations: spatial max pooling (extracting salient features) and spatial average pooling (capturing global contextual information) along

the channel dimension to achieve multi-granularity feature representation:

$$X_{Msam} = \max_{c \in C}(X_{sam}) \in \mathbb{R}^{B \times 1 \times H \times W} \tag{13}$$

$$X_{Asam} = \frac{1}{C}\left(\sum_{c=1}^{C} X_{sam}\right) \in \mathbb{R}^{B \times 1 \times H \times W} \tag{14}$$

Given the SAM2 encoder's layer-wise output features $X_{sam} \in \mathbb{R}^{B \times C \times H \times W}$, we first perform dual-path channel compression: spatial max pooling extracts salient features while spatial average pooling captures global context. These are then concatenated along the channel dimension:

$$X_{cat} = \text{Concat}(X_{Msam},\, X_{Asam}) \in \mathbb{R}^{B \times 2 \times H \times W} \tag{15}$$

To strengthen the SAM2 encoder's pixel-level positional semantic modeling, we first construct global feature representations via spatial average pooling and perform channel-wise contrastive enhancement with X_{cat}:

$$X_{global} = \text{AvgPool}(X_{cat}) \odot X_{cat} \tag{16}$$

A gated spatial attention (GSA) mechanism then models long-range spatial dependencies using 7×7 convolutional kernels:

$$X_{gsa} = X_{global} \odot \sigma\left(\text{Conv2D}_{7\times7,\,\text{padding}=3}(X_{global})\right) \tag{17}$$

The spatial attention weights generated by basic convolution and Sigmoid activation adaptively recalibrate SAM2 features:

$$X'_{sam} = X_{sam} \odot \sigma\left(\text{BasicConv}_{7\times7}(X_{gsa})\right) \tag{18}$$

4 Experiments and Results

Automated Cardiac Diagnosis Challenge The ACDC dataset originates from the 2017 MICCAI challenge of the same name, comprising cardiac MRI short-axis sequences from 100 patients collected by multiple French clinical centers, including the University Hospital of Dijon. This dataset encompasses five cardiac pathologies and normal cases: dilated cardiomyopathy (DCM) characterized by left ventricular enlargement, hypertrophic cardiomyopathy (HCM) marked by abnormal thickening of left ventricular myocardium, myocardial infarction (MINF) presenting left ventricular myocardial scarring, right ventricular dysfunction (RV-abnormal) manifesting structural/contractile abnormalities in the right ventricle, and normal cardiac anatomy. Pathological features primarily localize in the left ventricle (DCM, HCM, MINF) or right ventricle (RV-abnormal), requiring comprehensive multi-level cardiac segmentation and functional analysis through MRI short-axis views for accurate diagnosis.

4.1 Implementation Details

The experimental setup was established on an Ubuntu 23.10 operating system, utilizing Python 3.12.0 with PyTorch 1.10 deep learning framework accelerated by CUDA 12.1. The hardware configuration consisted of an NVIDIA A800-SXM4-80GB GPU and an Intel Xeon Platinum 8462Y+ CPU. We employed the preprocessed ACDC dataset for 2D medical image segmentation tasks. The SAMba-UNet model underwent 10,000 training iterations with a batch size of 12. Stochastic Gradient Descent (SGD) optimizer [2] was implemented with an initial learning rate of 0.01, momentum of 0.9, and weight decay coefficient of 0.0001. Model evaluation on the validation set was performed every 200 iterations, accompanied by checkpoint saving of optimal parameters.

4.2 Evaluation Metrics

The performance evaluation of SAMba-UNet and baseline methods employed a comprehensive set of metrics. The Dice coefficient was adopted to assess the overlap between predicted segmentation and ground truth labels. Intersection over Union (IoU) quantifies the ratio of overlapping area to the total union area. Accuracy measured the proportion of correctly classified pixels, while

Precision reflected the percentage of true positives among predicted positives. Sensitivity (Recall) evaluated the identification capability of true positive pixels, and Specificity indicated the correct exclusion rate of true negative pixels. Boundary matching was assessed through two metrics: The 95th percentile Hausdorff Distance (HD95), calculated as the 95th percentile of maximum distances between predicted and actual boundaries to mitigate outlier effects, and the Average Surface Distance (ASD) computed as the mean minimum distance between corresponding boundary points. Lower HD95 and ASD values indicate better performance, whereas higher values are preferred for all other metrics.

4.3 Quantitative Comparison

As demonstrated in the quantitative analysis of Table 1, SAMba-UNet achieves superior segmentation performance on the ACDC dataset. With an mDice score of 0.9103, it surpasses all comparative models by 0.71 percentage points over the suboptimal LeViT-UNet-384. Notably, the model exhibits performance gains of 0.94% and 0.84% for segmenting morphologically complex structures — the right ventricle (RV, 0.9039) and myocardium (MYO, 0.8935), respectively — which substantiates the effectiveness of its multi-scale feature modeling. Compared to conventional convolutional architectures (UNet++), attention-based methods (R50 Attn-UNet), and pure Transformer models (SwinUNet, UNETR), SAMba-UNet maintains comparable left ventricle (LV, 0.9335) segmentation accuracy (difference <2%) with mainstream approaches while achieving breakthrough collaborative segmentation performance through the novel integration of SAM2 architecture and state space models.

Table 1: Performance of Different Models on the ACDC Dataset

Model	mDice↑	RV	MYO	LV
UNetRonneberger et al. [2015]	0.8993	0.8682	0.8776	**0.9521**
UNet++Zhou et al. [2019]	0.8994	0.8730	0.8740	0.9507
R50 Attn-UNetOktay et al. [2018]	0.8675	0.8758	0.7920	0.9347
VIT-CUPChen et al. [2021]	0.8145	0.8146	0.7071	0.9218
R50 ViT-CUPChen et al. [2021]	0.8757	0.8607	0.8188	0.9475
TransUNetChen et al. [2021]	0.8923	0.8591	0.8667	<u>0.9511</u>
SwinUNetCao et al. [2022]	0.8928	0.8701	0.8637	0.9449
UNETRHatamizadeh et al. [2022]	0.8861	0.8529	0.8652	0.9402
LeViT-UNet-384Xu et al. [2023]	<u>0.9032</u>	<u>0.8955</u>	0.8764	0.9376
Mamba-UNetWang et al. [2024]	0.8997	0.8817	<u>0.8860</u>	0.9313
SAMba-UNet(Our)	**0.9103**	**0.9039**	**0.8935**	0.9335

5 Ablation Study

We conducted experiments to explore the following two aspects: (1)The effectiveness of different components of the model architecture; (2)Investigation of Effectiveness in Different Adapters and Their Components.

5.1 The effectiveness of different components of the model architecture

As shown in the ablation study results in Table 2, the complete model architecture (ALL) achieves optimal performance across all evaluation metrics, fully demonstrating the importance of synergistic interactions between modules. Further analysis reveals that removing critical components (e.g., the OCA module) leads to significant degradation in segmentation accuracy and weakened boundary localization capability, underscoring its critical role in the system. While the removal of other components (such as IRMLP and AdaptAttn) does not completely compromise basic model functionality, all exhibit varying degrees of negative impacts on core performance metrics. These observations effectively validate the efficient design of the overall architecture and the functional complementarity among different modules.

Table 2: The effectiveness of different components of the model architecture.

Configuration	mDice↑	mIoU↑	Acc↑	Pre↑	Sen↑	Spe↑	mHD95↓	ASD↓
w/o IRMLP	0.9023	0.8307	0.9978	0.8994	0.9116	<u>0.9989</u>	1.1841	0.4341
w/o AdaptAttn	0.9036	0.8314	0.9978	0.9014	0.9126	<u>0.9989</u>	1.8263	0.5154
w/o GCAA	0.9048	0.8341	0.9978	0.9016	0.9142	<u>0.9989</u>	1.6602	0.4461
w/o OCA	0.9037	0.8315	0.9978	0.8991	<u>0.9162</u>	<u>0.9989</u>	1.2468	0.3501
w/o BSEA	<u>0.9064</u>	<u>0.8357</u>	<u>0.9979</u>	<u>0.9055</u>	0.9128	0.9988	<u>1.1441</u>	<u>0.2947</u>
ALL	**0.9103**	**0.8392**	**0.9981**	**0.9157**	**0.9174**	**0.9991**	**1.0859**	**0.2611**

5.2 Investigation of Effectiveness in Different Adapters and Their Components

As shown in the ablation study results of Table 3, the complete adapter architecture (ALL) achieves optimal performance across all evaluation metrics, fully demonstrating the effectiveness of the collaborative working mechanism among components. Removing core modules (such as the Channel Attention or Refiner) leads to significant performance degradation, particularly evident in boundary localization accuracy and segmentation consistency metrics. Although the absence of the MLP-Adapter has a relatively minor impact on overall performance, it still causes precision loss in detailed features, indicating the module's irreplaceable role in feature refinement. Experimental results demonstrate that each submodule specifically enhances different capability dimensions of the model (e.g., region recognition accuracy, edge sharpness), while systematic integration of these purposefully designed components constitutes the key factor in achieving optimal segmentation performance.

Table 3: Investigation of Effectiveness in Different Adapters and Their Components.

Adapter	mDice↑	mIoU↑	Acc↑	Pre↑	Sen↑	Spe↑	mHD95↓	ASD↓
w/o ChannelAttn	0.9043	0.8324	0.9978	0.9041	0.9118	0.9990	1.1462	0.3189
w/o Refiner	0.9028	0.8307	0.9978	0.8967	0.9155	0.9989	1.9504	0.4936
w/o MLP-Adapter	0.9054	0.8331	0.9978	0.9043	0.9132	0.9990	1.1393	0.2881
ALL	**0.9103**	**0.8392**	**0.9981**	**0.9157**	**0.9174**	**0.9991**	**1.0859**	**0.2611**

6 Conclusion

The proposed SAMba-UNet innovatively integrates SAM2, Mamba, and UNet architectures, successfully addressing domain adaptation issues and fine-grained feature extraction challenges in medical image segmentation. Experimental results demonstrate that the Dynamic Feature Fusion Refiner effectively mitigates semantic discrepancies between natural and medical images, while the Heterogeneous Omni-Attention Convergence Module (HOACM) significantly enhances the collaborative modeling of global context and local details. On the ACDC dataset, the model outperforms existing methods in boundary segmentation accuracy for key cardiac structures (e.g., myocardium and ventricles) and sensitivity to complex pathologies.

References

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

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.

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.

Tapas Kumar Dutta, Snehashis Majhi, Deepak Ranjan Nayak, and Debesh Jha. Sam-mamba: Mamba guided sam architecture for generalized zero-shot polyp segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4655–4664. IEEE, 2025.

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

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. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.

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 *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Sheng He, Rina Bao, Jingpeng Li, Patricia Ellen Grant, and Yangming Ou. Accuracy of segment-anything model (sam) in medical image segmentation tasks. *CoRR*, 2023.

Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. Foundation model for advancing healthcare: challenges, opportunities and future directions. *IEEE Reviews in Biomedical Engineering*, 2024.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, and Ying Fu. Missformer: An effective transformer for 2d medical image segmentation. *IEEE transactions on medical imaging*, 42(5): 1484–1494, 2022.

Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024.

Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2025.

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

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024a.

Zhengyi Liu, Longzhen Wang, Xianyong Fang, Zhengzheng Tu, and Linbo Wang. Lfsamba: Marry sam with mamba for light field salient object detection. *IEEE Signal Processing Letters*, 2024b.

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

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H Maier-Hein. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv preprint arXiv:2304.05396*, 2023.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International conference on machine learning*, pages 29441–29454. PMLR, 2023.

Saheed Sanyaolu. Integration of machine learning in imaging analysis for clinical diagnosis of cardiovascular diseases. *Cardiology*, 2:100006, 2025.

Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24*, pages 36–46. Springer, 2021.

Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*, 2024.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

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

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.

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)*, pages 42–53. Springer, 2023.

Shu Yang, Yihui Wang, and Hao Chen. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 296–306. Springer, 2024.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. Mitigate position bias in large language models via scaling a single dimension. *arXiv preprint arXiv:2406.02536*, 2024.

Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24*, pages 14–24. Springer, 2021.

Xiao-Yun Zhou, Jian-Qing Zheng, Peichao Li, and Guang-Zhong Yang. Acnn: a full resolution dcnn for medical image segmentation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8455–8461. IEEE, 2020.

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.

Wenqiang Zu, Shenghao Xie, Qing Zhao, Guoqi Li, and Lei Ma. Embedded prompt tuning: Towards enhanced calibration of pretrained models for medical images. *Medical Image Analysis*, 97:103258, 2024.