Collaborative Feature and Persona Enhancement for Trustworthy Medical Foundation Models

Fanqi Kong
Zhejiang University
Guangxi University
fanqikong@st.gxu.edu.cn

Ziming Zhao*
Zhejiang University
zhaoziming@zju.edu.cn

Abstract

Visual state-space models (SSMs) such as Mamba have emerged as strong backbones for medical image segmentation due to their ability to capture long-range dependencies with linear computational complexity. However, existing Mamba-based architectures provide limited support for explicit feature selection, targeted feature enhancement, and dedicated multi-scale representation learning, leaving them vulnerable to confusing anatomical structures and imaging noise. Moreover, segmentation models deployed in real-world clinical environments must remain robust across heterogeneous demographic profiles without amplifying spurious or stereotype-like correlations. We introduce *Persona-guided Collaborative Enhancement and Inception GroupMamba UNet (Persona-Guided CEIGM-UNet)*, a Mamba-based segmentation framework that addresses these limitations. Built upon a GroupMamba encoder, our design incorporates: (i) a *Collaborative Feature Enhancement Layer* (CFEL) that integrates attention-guided refinement, dynamic up-convolution, and multi-scale enhancement gating; (ii) a *Modulated Inception Group Mamba Layer* (MIGML) that couples multi-scale local pattern extraction with long-range dependency modeling; and (iii) a lightweight *Demographic-Aware Persona Modulation* (DAPM) branch that maps demographic meta-information to bounded channel-wise modulation factors, enabling mild, controlled feature adaptation. Experiments on the Synapse and ACDC datasets show that the CEIGM-UNet backbone achieves state-of-the-art performance with fewer parameters and competitive FLOPs. A preliminary fairness evaluation on Synapse, assessing equalized odds differences and generalized Dice disparities across age and sex, suggests that persona-guided modulation can reduce group-wise performance gaps relative to strong Transformer baselines.

1 Introduction

Medical image segmentation underpins computer-aided diagnosis and image-guided therapy by delineating organs and lesions in CT or MRI scans [1, 2]. Convolutional neural networks (CNNs), exemplified by U-Net, remain the dominant paradigm for this task, yet their inherently local receptive fields restrict their ability to capture long-range anatomical dependencies [3, 1, 4]. Vision Transformers address this limitation through global self-attention, but their quadratic complexity with respect to sequence length leads to substantial memory and computational overhead [5, 6, 7, 8].

State space models (SSMs), such as S4 and the recently proposed Mamba, offer a compelling alternative by modeling long-range dependencies with linear-time complexity and have shown promise in visual and medical segmentation tasks [9, 10, 11]. Nevertheless, current Mamba-based designs largely focus on global sequence modeling and provide limited support for explicit feature selection

*Corresponding author

or structured multi-scale feature learning, particularly within decoder stages. This omission renders them vulnerable to confounding background structures and to the challenges posed by small or irregular anatomical regions.

Beyond accuracy, segmentation models deployed in heterogeneous clinical environments must also remain reliable and equitable across patient subpopulations that differ in age, sex, and other demographic attributes. In other words, a socially responsible model should sustain consistent performance across groups while avoiding over-reliance on spurious demographic correlations [12, 13, 14]. However, existing Mamba-based architectures do not incorporate explicit mechanisms for analyzing, mitigating, or controlling demographic-conditioned behavior, limiting their applicability in fairness-critical clinical settings.

To address these limitations, we intend to present a segmentation architecture that (i) explicitly enhances and propagates target-relevant features across scales, (ii) integrates multi-scale local modeling with the long-range dependency strengths of state-space models, and (iii) enables controlled demographic-aware conditioning to support responsible deployment in heterogeneous clinical populations. Motivated by these goals, we design a Mamba-based framework that jointly improves feature discrimination, representation granularity, and fairness-oriented adaptability.

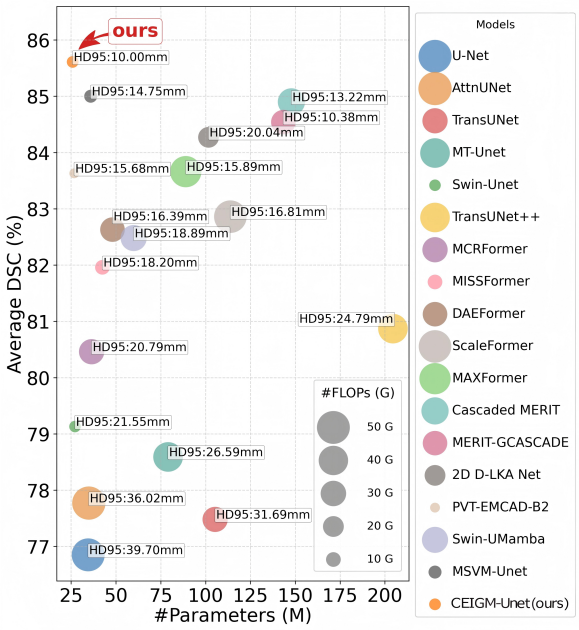

Figure 1: Comparison of different methods on the Synapse dataset in terms of model complexity and segmentation performance. The proposed CEIGM-UNet achieves the highest average DSC and the lowest HD95 with the fewest parameters.

In summary, this paper makes the following contributions.

- We propose *CEIGM-UNet*, a new Mamba-based U-shaped architecture, which integrates a Collaborative Feature Enhancement Layer (CFEL) with a Modulated Inception Group Mamba Layer (MIGML). CFEL explicitly selects, enhances, and propagates target-relevant features, while MIGML couples multi-scale local pattern extraction with long-range dependency modeling.
- We introduce a lightweight Demographic-Aware Persona Modulation (DAPM) branch that conditions intermediate feature channels on a low-dimensional persona vector derived from demographic meta-information. A bounded modulation budget restricts the magnitude of channel re-weighting, encouraging calibrated adaptation rather than stereotype amplification.
- We perform extensive experiments and fairness analysis. On the Synapse abdominal CT and ACDC cardiac MRI datasets, CEIGM-UNet achieves state-of-the-art segmentation performance with fewer parameters and competitive FLOPs. A preliminary fairness evaluation across age and sex on Synapse shows that enabling DAPM reduces group-wise disparities without degrading overall performance.

2 Related Work

2.1 CNN-Based and Transformer-Based Segmentation

U-Net and its numerous variants form the foundation of modern medical image segmentation, leveraging hierarchical convolutional encoders and decoders to capture spatial structure [1, 15]. While effective, their inherently local receptive fields limit their ability to model long-range anatomical

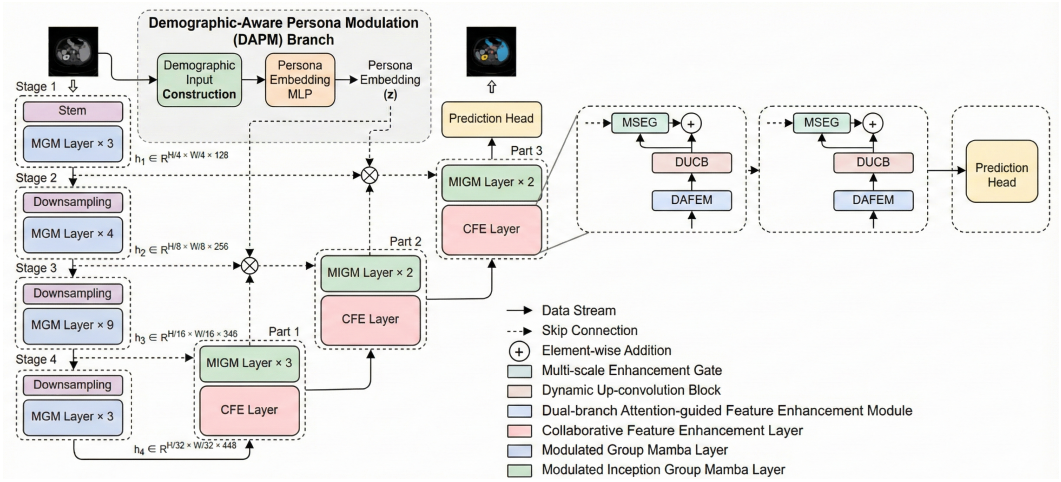

Figure 2: Overall architecture of the proposed Persona-Guided CEIGM-UNet. The network consists of a GroupMamba-T encoder, a CFEL-augmented decoder, a prediction head, and a demographic-aware persona modulation (DAPM) branch.

dependencies. Transformer-based architectures address this limitation by incorporating global self-attention, and are often combined with CNNs to balance local detail with global context. However, the quadratic complexity of self-attention with respect to sequence length renders such models computationally expensive for high-resolution medical imaging [5, 6, 7, 8, 16].

2.2 Mamba-Based Segmentation

State space models (SSMs), including S4 and the recently introduced Mamba, provide a linear-time alternative for modeling long-range dependencies [9]. Their strong expressiveness and efficiency have motivated the development of visual Mamba variants in hybrid CNN–SSM frameworks and in pure SSM-based encoders [10, 17, 11]. Despite their promise, existing Mamba-based designs largely emphasize global sequence modeling and offer limited mechanisms for explicit feature selection, multi-scale feature fusion, or structured decoder design, which restricts their ability to handle fine-grained targets and small, irregular anatomical structures.

2.3 Demographic-Aware and Persona-Guided Modeling

Fairness-aware medical imaging research has increasingly focused on assessing and mitigating performance disparities across demographic groups through subgroup evaluation, bias analysis, and regularization strategies [12, 18, 19, 20]. In parallel, persona-guided conditioning has been explored in vision-language and foundation models, where low-dimensional vectors modulate internal representations to influence downstream behavior [21, 22]. Inspired by these developments, we extend Mamba-based segmentation by incorporating a lightweight, explicitly bounded persona modulation branch and investigate its effect on subgroup fairness without compromising overall segmentation performance.

3 Method

3.1 Overview of Persona-Guided CEIGM-UNet

As illustrated in Figure 2, Persona-Guided CEIGM-UNet adopts a U-shaped segmentation architecture composed of a GroupMamba-based encoder, a CFEL-enhanced decoder, and standard skip connections. The network begins with a stem block for initial feature extraction, after which the encoder processes the input through four hierarchical stages built upon GroupMamba-T. Each stage contains several Modulated Group Mamba (MGM) Layers, with downsampling applied at the beginning of all but the first stage.

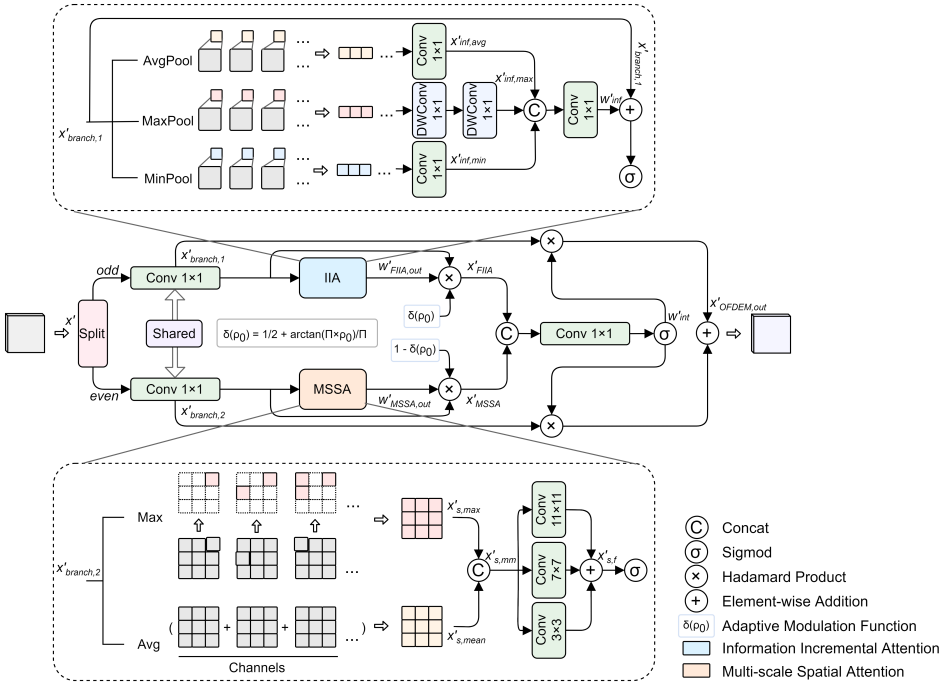

Figure 3: Structure of the Dual-branch Attention-guided Feature Enhancement Module (DAFEM), which contains an information-incremental attention branch and a multi-scale spatial attention branch that collaboratively enhance target-related features.

The decoder mirrors this hierarchical structure and is organized into three levels. Each level first applies a Collaborative Feature Enhancement Layer (CFEL) to selectively amplify target-related information and propagate it to higher spatial resolutions, followed by multiple Modulated Inception Group Mamba Layers (MIGML) that jointly refine and fuse multi-scale representations. A lightweight attention module coupled with a final convolutional layer forms the prediction head responsible for generating voxel-wise segmentation logits.

To enable demographic-aware conditioning, a separate Demographic-Aware Persona Modulation (DAPM) branch maps a low-dimensional persona vector, constructed from demographic meta-information, to stage-wise, channel-wise modulation factors. These factors mildly adjust intermediate feature maps in a bounded manner, allowing controlled adaptation. When the modulation strength is set to zero, the model reduces to the plain CEIGM-UNet backbone, enabling evaluation in settings without demographic annotations.

3.2 Collaborative Feature Enhancement Layer

The Collaborative Feature Enhancement Layer (CFEL) is designed to strengthen the selection, refinement, and propagation of target-relevant information by jointly enhancing deep semantic features and recalibrating shallow spatial features. Given a deep feature map $\mathbf{F}_{\text{deep}}$ and a shallow feature map $\mathbf{F}_{\text{shallow}}$, CFEL first enhances the deep representation, then performs content-adaptive upsampling, and finally integrates the resulting features with recalibrated shallow features.

The process begins with the Dual-branch Attention-guided Feature Enhancement Module (DAFEM), which aggregates multi-branch pooling statistics to estimate the amount of target-related information present in each channel and fuses these statistics with multi-scale spatial attention. This produces an enhanced deep representation $\mathbf{F}_{\text{enh}}$ with irrelevant responses effectively suppressed; its internal structure is illustrated in Figure 3. The enhanced features are then passed to the Dynamic Up-convolution Block (DUCB), which performs content-adaptive upsampling by generating sampling offsets and applying lightweight convolutions (Figure 4). This dynamic resampling mech-

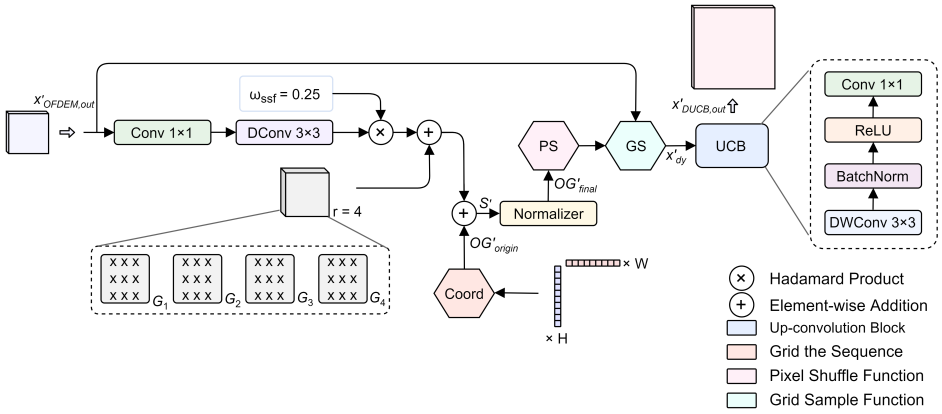

Figure 4: Dynamic up-convolution block (DUCB). It integrates the content-adaptive resampling idea of DySample with a lightweight convolutional design to perform efficient dynamic upsampling.

anism enables $\mathbf{F}_{\text{enh}}$ to be faithfully propagated to higher spatial resolutions, improving semantic consistency across scales.

To incorporate detailed spatial cues, CFEL further employs a Multi-scale Enhancement Gate (MSEG). Conditioned on the upsampled features $\mathbf{F}_{\text{up}}$, MSEG recalibrates the shallow feature map by producing fine-grained, multi-scale attention weights that emphasize boundaries and delicate anatomical structures while suppressing noise and background clutter, as depicted in Figure 5. The final output of CFEL is obtained by fusing the two refined streams:

$$\mathbf{F}_{\text{enh}} = \text{DAFEM}(\mathbf{F}_{\text{deep}}), \tag{1}$$

$$\mathbf{F}_{\text{up}} = \text{DUCB}(\mathbf{F}_{\text{enh}}), \tag{2}$$

$$\mathbf{F}_{\text{out}} = \mathbf{F}_{\text{up}} + \text{MSEG}(\mathbf{F}_{\text{shallow}}, \mathbf{F}_{\text{up}}). \tag{3}$$

Through this coarse-to-fine enhancement pipeline, CFEL provides a structured mechanism for selectively amplifying semantic information, propagating it across resolutions, and harmonizing it with spatially rich yet noisier shallow features, ultimately improving the segmentation of small, irregular, or low-contrast anatomical regions.

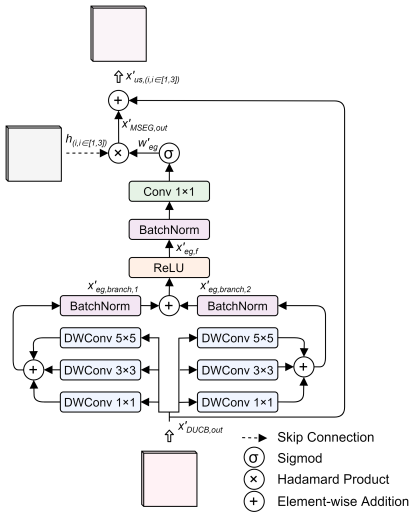

Figure 5: Multi-scale enhancement gate.

3.3 Modulated Inception Group Mamba Layer

While Mamba-based architectures excel at modeling long-range dependencies with linear complexity, they typically lack explicit mechanisms for structured multi-scale feature learning. To address this gap, we draw inspiration from Inception-style designs and redesign the feed-forward network (FFN) within the Modulated Group Mamba Layer (MGML), yielding the Modulated Inception Group Mamba Layer (MIGML).

Let $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ denote the input feature map. MIGML first applies a GroupMamba block to capture global and sequential dependencies

$$\mathbf{U} = \text{GroupMamba}(\mathbf{X}), \tag{4}$$

and refines representation via Inception Feed-Forward Network (IFFN) with a residual connection

$$\mathbf{V} = \text{IFFN}(\mathbf{U}) + \mathbf{U}, \tag{5}$$

followed by a second residual connection to preserve the original input

$$\mathbf{Y} = \mathbf{V} + \mathbf{X}. \tag{6}$$

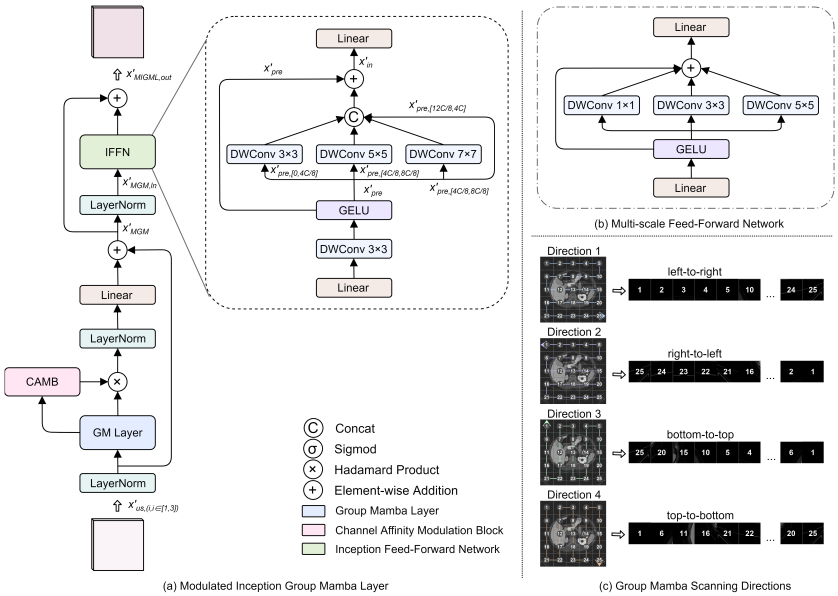

Figure 6: Architecture of the modulated inception group Mamba layer, where the Inception Feed-Forward Network replaces the original feed-forward network with a channel allocation strategy.

The IFFN begins by expanding the channel dimension and applying a depthwise convolution followed by GELU activation. The expanded channels are then partitioned into multiple groups, each processed by depthwise convolutions with kernel sizes 3×3, 5×5, and 7×7, along with an identity pathway. The multi-branch outputs are concatenated and projected back to the original channel dimension. This design allows MIGML to simultaneously capture fine, intermediate, and coarse spatial patterns while maintaining the efficiency and long-range modeling capability of the underlying Mamba block. The full architecture of MIGML is illustrated in Figure 6.

By embedding an Inception-style multi-scale operator within the MGML structure, MIGML provides a mechanism for enriching local details and enhancing feature granularity capabilities that are crucial for segmenting anatomically diverse organs and structures with varying spatial scales.

3.4 Demographic-Aware Persona Modulation

The DAPM branch introduces demographic-aware conditioning while explicitly limiting its strength.

Persona Vector Construction. For each case, we define a meta-information vector $\mathbf{m} \in \mathbb{R}^{d_m}$ that encodes demographic attributes such as age group and sex. Categorical variables are one-hot encoded, and continuous variables (*e.g.,* age) are normalized. The meta-information vector is mapped to a compact persona embedding using a two-layer MLP:

$$\mathbf{z} = f_{\text{persona}}(\mathbf{m}) \in \mathbb{R}^{d_z}. \tag{7}$$

On Synapse, we assume access to age and sex; age is grouped into two cohorts ($<$50 vs. $\geq$50 years), and sex is encoded as a binary one-hot vector.

Stage-Wise Channel Modulation. For the feature map at stage l, $\mathbf{F}^{(l)} \in \mathbb{R}^{C_l \times H_l \times W_l}$, DAPM generates a channel-wise modulation vector via a stage-specific MLP followed by normalization and a bounded nonlinearity:

$$\mathbf{a}^{(l)} = f^{(l)}_{\text{stage}}(\mathbf{z}), \tag{8}$$

$$\hat{\mathbf{a}}^{(l)} = \frac{\mathbf{a}^{(l)}}{\|\mathbf{a}^{(l)}\|_2 + \varepsilon}, \tag{9}$$

$$\boldsymbol{\gamma}^{(l)} = \mathbf{1} + \rho_0 \cdot \tanh(\hat{\mathbf{a}}^{(l)}), \tag{10}$$

Table 1: Comparison results on the Synapse dataset.

Method	Params (M)↓	FLOPs (G)↓	DSC (%)↑	HD95 (mm)↓
U-Net	34.53	50.22	76.85	39.70
TransUNet	105.32	29.35	77.48	31.69
Swin-UNet	27.17	6.14	79.13	21.55
MISSFormer	42.46	9.89	81.96	18.20
PVT-EMCAD-B2	26.76	**4.45**	83.63	15.68
Swin-UMamba	59.89	31.48	82.48	18.89
MSVM-UNet	35.93	7.80	85.00	14.75
CEIGM-UNet (Ours)	**25.86**	6.31	**85.61**	**10.00**

where $\rho_0 \in (0, 1)$ is the modulation budget. Each scaling factor therefore satisfies $\gamma_c^{(l)} \in [1-\rho_0, 1+\rho_0]$. The feature map is modulated channel-wise:

$$\tilde{\mathbf{F}}_{c,h,w}^{(l)} = \gamma_c^{(l)} \cdot \mathbf{F}_{c,h,w}^{(l)}. \tag{11}$$

We insert DAPM blocks at the output of selected encoder stages and decoder CFELs. When $\rho_0 = 0$, DAPM is disabled, and the model reduces to CEIGM-UNet.

To discourage unnecessary reliance on persona information, we optionally add a regularization term

$$\mathcal{L}_{\text{mod}} = \sum_l \alpha_l \|\boldsymbol{\gamma}^{(l)} - \mathbf{1}\|_2^2, \tag{12}$$

and optimize the total loss

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda \mathcal{L}_{\text{mod}}, \tag{13}$$

where $\mathcal{L}_{\text{seg}}$ is the segmentation loss and λ controls regularization strength.

4 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of our proposal.

4.1 Datasets

Synapse. The Synapse dataset from the MICCAI 2015 Multi-Atlas Abdominal Labeling Challenge contains 30 abdominal CT scans with eight organ annotations [23]. Following common practice, 18 volumes are used for training and 12 for validation. For fairness experiments, we assume access to age and sex metadata and form four demographic subgroups by crossing age cohort (< 50 vs. ≥ 50) and sex (male vs. female).

ACDC. The ACDC dataset consists of 100 cardiac MRI scans with annotations for left ventricle (LV), right ventricle (RV), and myocardium (MYO) [24]. We follow the standard split: 70 cases for training, 10 for validation, and 20 for testing.

4.2 Training Setup and Metrics

We implement Persona-Guided CEIGM-UNet in PyTorch and initialize the encoder with ImageNet-1k pretrained GroupMamba-T weights. Input slices are resized to 224×224 and augmented with flipping, rotation, Gaussian noise, and contrast adjustment. We use AdamW optimization, train for 250 epochs on Synapse and 300 epochs on ACDC, and adopt a hybrid loss consisting of cross-entropy and Dice loss:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Dice}}. \tag{14}$$

For the main segmentation results, we disable DAPM ($\rho_0 = 0$); For fairness experiments on Synapse, we enable DAPM with $\rho_0 = 0.1$.

We report Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95) as primary metrics, along with the number of parameters and FLOPs. For fairness, we compute equalized

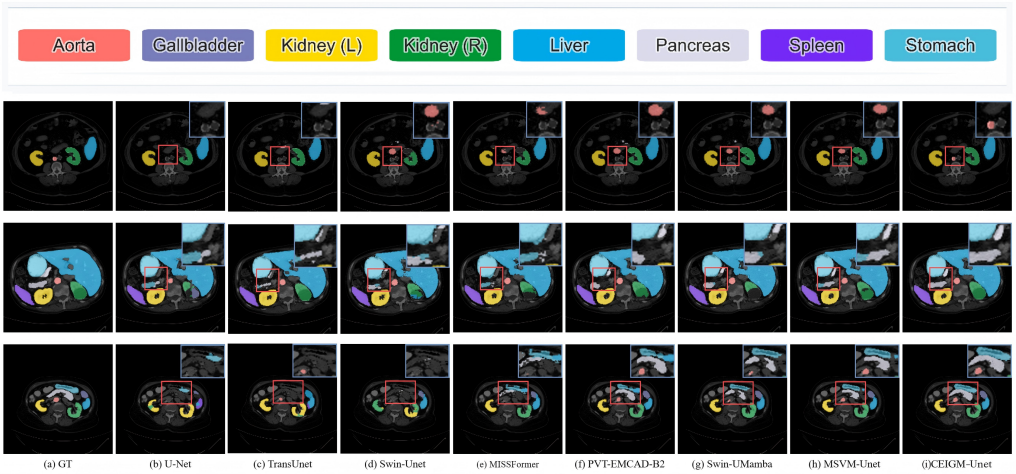

Figure 7: Qualitative comparison on the Synapse dataset. From left to right: GT, U-Net, TransUNet, Swin-UNet, MISSFormer, PVT-EMCAD-B2, Swin-UMamba, MSVM-UNet, and CEIGM-UNet.

odds difference (EOD) and generalized Dice disparity (GDD) across age-sex groups. EOD measures the maximum absolute difference in true positive rate and false positive rate between groups; GDD quantifies discrepancies in generalized Dice scores with organ-size weighting. Lower EOD and GDD indicate smaller disparities.

4.3 Results on Synapse

Table 1 compares CEIGM-UNet with representative CNN-, Transformer-, and Mamba-based segmentation models on the Synapse dataset. CEIGM-UNet achieves the highest average DSC and the lowest HD95 among all methods, while using fewer parameters than most baselines and maintaining competitive FLOPs comparable to lightweight designs such as PVT-EMCAD-B2. These results indicate that the combination of CFEL and MIGML substantially improves feature discrimination, spatial consistency, and multi-scale representation quality. Figure 7 presents qualitative comparisons across several methods. CEIGM-UNet produces sharper boundaries and more complete anatomical structures, particularly for small or morphologically irregular organs such as the gallbladder and pancreas. The model also better suppresses distracting background responses and reduces leakage into adjacent tissues, highlighting the effectiveness of the proposed feature enhancement and multi-scale modeling mechanisms in challenging abdominal regions.

4.4 Results on ACDC

Table 2 summarizes the performance of CEIGM-UNet on the ACDC cardiac MRI dataset. Our model achieves the highest DSC and the lowest HD95 across the LV, RV, and MYO classes, surpassing both Transformer-based and Mamba-based baselines. These results demonstrate that the proposed architecture generalizes effectively across different imaging modalities (CT vs. MRI) and anatomical contexts (abdominal vs. cardiac), despite being trained under the same unified framework. Qualitative comparisons in Figure 8 further illustrate these advantages. CEIGM-UNet preserves sharper myocardial boundaries, reduces prediction leakage near ventricular edges, and yields more spatially consistent delineations compared to prior methods.

Table 2: Comparison results on the ACDC dataset.

Method	DSC (%)↑	HD95 (mm)↓
TransUNet	89.71	2.01
Swin-UNet	90.00	2.62
Swin-UMamba	92.22	1.68
MSVM-UNet(Ours)	92.58	1.41
CEIGM-UNet	**92.81**	**1.05**

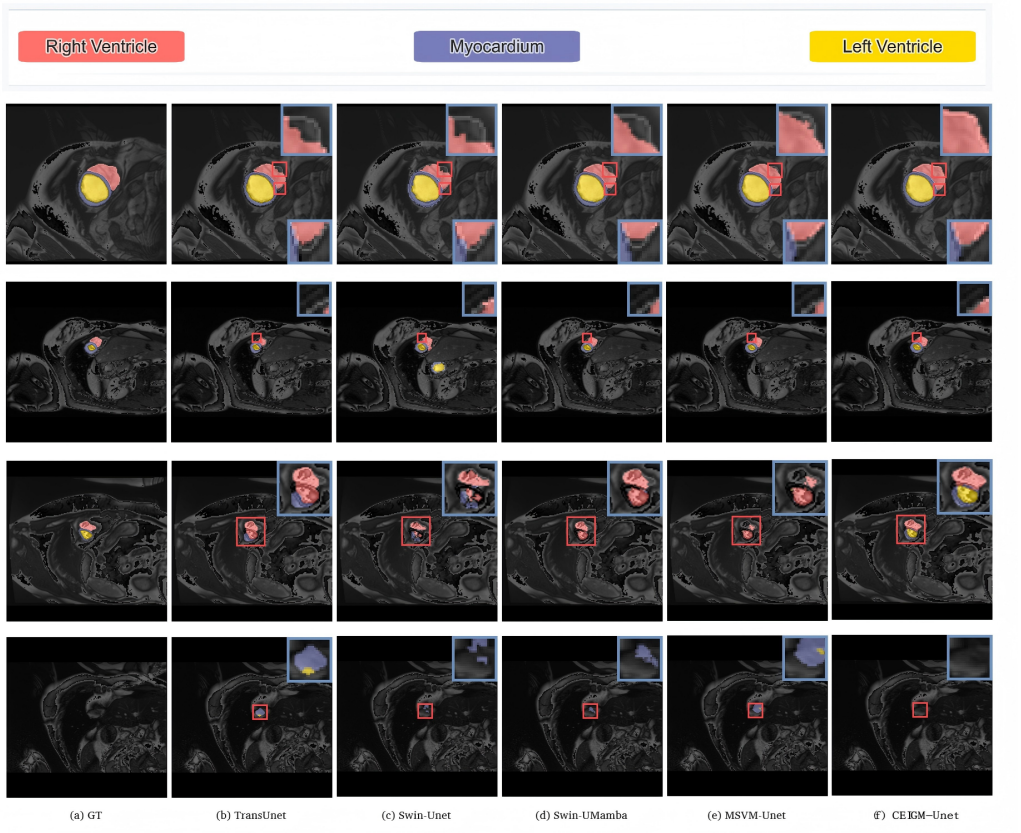

Figure 8: Qualitative comparison on the ACDC dataset. From left to right: GT, TransUNet, Swin-UNet, Swin-UMamba, MSVM-UNet, and CEIGM-UNet.

4.5 Fairness Evaluation Across Demographic Groups

We assess fairness on the Synapse dataset by measuring equalized odds difference (EOD) and generalized Dice disparity (GDD) across four age-sex subgroups. TransUNet and Swin-UNet serve as baselines that do not incorporate demographic information. For persona-guided CEIGM-UNet, we activate the DAPM branch with a modest modulation budget of $\rho_0 = 0.1$. As shown in Table 3, persona-guided CEIGM-UNet achieves the lowest disparities on both metrics while simultaneously maintaining the highest DSC (Table 1). These results indicate that lightly bounded persona modulation can reduce subgroup performance gaps without degrading accuracy, suggesting a viable pathway for integrating demographic awareness into segmentation models without amplifying bias.

Table 3: Fairness evaluation.

Model	EOD	GDD
TransUNet	0.074	0.056
Swin-UNet	0.059	0.048
Per. CEIGM-UNet	**0.031**	**0.029**

5 Conclusion

In this paper, we introduce persona-Guided CEIGM-UNet, a Mamba-based U-shaped segmentation network that integrates a series of tailor-made components, *i.e.,* CFEL, MIGML, and DAPM. Across Synapse and ACDC, CEIGM-UNet achieves state-of-the-art performance with competitive efficiency. The fairness analysis on Synapse further shows that modest Persona-Guided modulation can reduce age-sex disparities without compromising overall accuracy. Extending demographic awareness to richer attributes, broader cohorts, and additional imaging modalities represents a promising direction for advancing socially responsible medical foundation models.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[2] Neeraj Kumar, Rakesh Verma, et al. Medical image segmentation: A decade of progress and future directions. *IEEE Transactions on Medical Imaging*, 39(8):2596–2630, 2020.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440. IEEE, 2015.

[4] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021.

[6] Jin Chen, Yong Luo, Shujun Yan, Yu Zhou, Xue Yang, Meng Yao, et al. Transunet: Transformers make strong encoders for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

[7] 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 *Computer Vision – ECCV 2022 Workshops*, volume 13803 of *Lecture Notes in Computer Science*, pages 205–218. Springer, 2023.

[8] Xiyue Huang et al. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(5):1484–1494, 2023.

[9] Shitong Gu, Pedro Chaves, Siddhartha Goyal, Lester Mackey, Vedant Sahajpal, Barret Zoph, and Zihang Dai. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2402.13606*, 2024.

[10] Amr Shaker et al. Groupmamba: Efficient group-based visual state space model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

[11] Chaowei Chen, Li Yu, Shiquan Min, and Shunfang Wang. MSVM-UNet: Multi-scale vision mamba UNet for medical image segmentation. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024.

[12] Liang Xu, Sarah Lee, Samuel L. Baxter, et al. Algorithmic fairness in computational medicine. *Journal of the American Medical Informatics Association*, 29:2013–2024, 2022.

[13] Karen Drukker, Wenjun Chen, Judy Gichoya, et al. Toward fairness in artificial intelligence for medical image analysis: Identification and mitigation of potential biases in the roadmap from data collection to model deployment. *Journal of Medical Imaging*, 10:061104, 2023.

[14] Daiju Ueda, Taichi Kakinuma, Shohei Fujita, Koji Kamagata, Yasutaka Fushimi, Rintaro Ito, and Yusuke Fushimi. Fairness of artificial intelligence in healthcare: Review and recommendations. *Japanese Journal of Radiology*, 2024.

[15] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias Heinrich, Kensaku Mori, Steven McDonagh, Nils Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. In *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.

[16] Md Mostafijur Rahman and Radu Marculescu. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In *Proceedings of the 6th International Conference on Medical Imaging with Deep Learning (MIDL)*, 2023.

[17] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-UMamba: Mamba-based UNet with imagenet-based pretraining. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 15009 of *Lecture Notes in Computer Science*, pages 615–625. Springer, 2024.

[18] Dilermando Queiroz, Anderson Carlos, André Anjos, and Lilian Berton. Fair foundation models for medical image analysis: Challenges and perspectives. *arXiv preprint arXiv:2502.16841*, 2025.

[19] Ruinan Jin, Zikang Xu, Yuan Zhong, Qiongsong Yao, Qi Dou, S. Kevin Zhou, and Xiaoxiao Li. Fairmedfm: Fairness benchmarking for medical imaging foundation models. *arXiv preprint arXiv:2407.00983*, 2024.

[20] Andy Liu, Mona Diab, and Daniel Fried. Evaluating large language model biases in persona-steered generation. *arXiv preprint arXiv:2405.20253*, 2024.

[21] Matthew Caron and Rahul Srivastava. Better prompting techniques for personality expression in language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[22] Shiyue Li, Michael Zeng, et al. Personality editing and control in neural text generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[23] MICCAI Multi-Atlas Abdomen Labeling Challenge Organizers. Segmentation of the abdominal organs in CT images for clinical applications. In *MICCAI Multi-Atlas Labeling Beyond the Cranial Vault (BTCV) Workshop and Challenge*, 2015.

[24] Olivier Bernard et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.