Collaborative Feature and Persona Enhancement for Trustworthy Medical Foundation Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Foundation models promise to democratize access to high-quality medical decision support by learning from vast quantities of data, but unmitigated biases in the data and model architecture can undermine their trustworthiness. Inspired by recent advances in persona-steered language modelling and efficient vision transformers, we propose a new architecture that jointly learns fine-grained medical image representations and patient personas while accounting for fairness and cognitive plausibility. Our model builds upon a multi-scale U-shaped backbone with collaborative feature enhancement and Group Mamba layers. We introduce a persona module that conditions intermediate features on demographic embeddings and a psychologically motivated modulation function. Experiments on multi-organ CT (Synapse) and cardiac MR (ACDC) benchmarks demonstrate competitive segmentation accuracy with substantially fewer parameters than conventional transformers. We further evaluate persona steerability and bias, showing that our approach produces more authentic persona behaviors than baseline methods while maintaining equitable performance across demographic groups. Finally, we discuss psychological foundations and ethical considerations of persona-aware medical foundation models and outline directions for responsibly developing trustworthy AI in healthcare.

1 Introduction

The past decade has witnessed tremendous progress in representation learning across natural language processing and computer vision, culminating in "foundation models" that transfer knowledge across diverse downstream tasks. Seminal works such as fully convolutional networks for semantic segmentation[Long et al., 2015], U-Net for biomedical image segmentation[Ronneberger et al., 2015], residual neural networks[He et al., 2016], vision transformers[Dosovitskiy et al., 2021] and subsequent hybrid architectures like TransUNet and Swin-UNet[Chen et al., 2021, Cao et al., 2021] have laid the groundwork for modern segmentation systems. In parallel, research on volumetric networks such as V-Net[Milletari et al., 2016] and robust semantic decoders like DeepLab[Chen et al., 2018] has demonstrated the benefits of multi-scale feature aggregation and long-range context. More recently, state space sequence models such as Mamba[Gu et al., 2024] and efficient attention variants have begun to replace transformers, providing linear time complexity and strong inductive biases for structured data.

In medical imaging, foundation models promise to democratize access to high-quality diagnostics by learning from vast corpora of computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound scans. However, naive deployment of these models risks perpetuating biases present in training data and algorithmic design. Reviews on algorithmic fairness in computational medicine highlight that many existing systems perform worse on minority populations and may reinforce health disparities[Xu et al., 2022, Koçak et al., 2025]. Interdisciplinary surveys further argue that

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

fairness cannot be achieved by post-hoc calibration alone: instead, interventions should span data curation, model architecture and deployment guidelines[Queiroz et al., 2025, Mehrabi et al., 2022]. Notably, Hardt et al. formalized the concept of equality of opportunity and equalized odds in supervised learning, providing statistical definitions of fairness that can be measured during model evaluation[Hardt et al., 2016]. Recent benchmarks such as FairMedFM integrate dozens of datasets to systematically assess the fairness of medical imaging foundation models and reveal persistent disparities across sensitive attributes[Jin et al., 2024].

Another emerging theme is persona steering, conditioning models on user profiles or demographic embeddings to produce tailored outputs. Early dialogue systems incorporated persona descriptors to improve coherence, but large language models (LLMs) have revealed subtle biases in such conditioning. Liu et al. show that reinforcement-learning-from-human-feedback (RLHF) reduces steerability toward incongruous personas and decreases response diversity[Liu et al., 2024]. Dash et al. find that persona-assigned LLMs exhibit human-like motivated reasoning, selectively aligning with identity-congruent statements and lowering veracity discernment by nearly ten percent[Dash et al., 2025]. To probe the psychometric validity of LLMs, Jiang et al. administer Big Five personality tests to LLM personas and observe that generated texts convey discernible personality traits, albeit with reduced authenticity when annotators know the content is AI[Jiang et al., 2024]. Surveys of bias and fairness in large language models catalogue representation gaps, toxicity and stereotype propagation across demographic groups, underscoring the need for responsible persona control[Gallegos et al., 2023]. Outside language, persona conditioning has rarely been explored for vision tasks, particularly in medical domains where patient demographics strongly influence disease presentation and treatment.

This work connects these strands by introducing a persona-aware medical segmentation framework. Our starting point is the CEIGM-UNet, an efficient U-Net variant that couples a collaborative feature enhancement layer (CFEL) with modulated Group Mamba (MGM) modules for multi-scale representation learning. Inspired by cognitive theories of identity and perception, we propose a novel modulation function that injects demographic embeddings into the network while saturating their influence at extreme values. We further develop metrics to evaluate persona authenticity, steerability and fairness in the context of image segmentation. Through experiments on multi-organ CT and cardiac MR benchmarks, we demonstrate that our persona-aware CEIGM-UNet achieves competitive segmentation accuracy with significantly fewer parameters than transformer-based models while improving fairness across demographic groups.

Our contributions are threefold. First, we extend CEIGM-UNet with a persona conditioning module that enables end-to-end learning of medical image segmentation and user-aware steering. Second, we synthesize psychological and algorithmic insights to derive an adaptive modulation function and develop evaluation metrics grounded in equalized odds and generalized Dice disparity. Third, we provide an extensive review of related work on segmentation architectures, fairness in medical AI, and persona modelling, laying the foundation for socially responsible and trustworthy medical foundation models.

2 Related Work

This section situates our contribution within three bodies of literature: persona modelling in language models, fairness in medical foundation models, and medical image segmentation networks.

2.1 Persona modelling and bias in large language models

Research on persona modelling originated in conversational agents, where conditioning on a speaker profile improves coherence and engagement. Subsequent works explored controllable generation via prompt engineering and reinforcement learning. However, as language models scale, persona conditioning can amplify biases and reduce diversity. Liu et al. systematically probe LLMs using congruous and incongruous personas and observe that RLHF-tuned models are 9.7% less steerable toward incongruous personas and produce more stereotypical outputs than raw models[Liu et al., 2024]. Dash et al. assign political personas to LLMs and find that persona-assigned models exhibit human-like motivated reasoning and lower veracity discernment of misinformation[Dash et al., 2025]. Jiang et al. introduce PersonaLLM, evaluating LLMs on the Big Five Inventory and story writing tasks; they show that LLM personas manifest consistent personality traits but that humans perceive them as less authentic when aware of AI authorship[Jiang et al., 2024]. Surveys of bias and

fairness in large language models catalog representation gaps, toxicity and stereotype propagation across demographic groups, advocating for systematic bias auditing and mitigation[Gallegos et al., 2023]. Complementary studies propose better prompting techniques, persona editing and life-story construction to steer LLMs toward desired personas while preserving truthfulness and safety[Caron and Srivastava, 2022, Li et al., 2023b, Park et al., 2023].

2.2 Fairness and trustworthiness in medical foundation models

Ensuring equitable performance across demographic groups is critical for clinical deployment. Algorithmic fairness in computational medicine surveys the literature on bias sources, fairness metrics and mitigation strategies, emphasising the need for domain-specific evaluations and multi-stakeholder collaboration[Xu et al., 2022]. Bias manifests not only in data but also in annotation practices, algorithm design and deployment contexts[Koçak et al., 2025]. Mehrabi et al. provide a comprehensive taxonomy of fairness definitions and categorize sources of bias across data, modelling and evaluation, highlighting that many definitions (e.g., demographic parity, equalized odds, predictive parity) can be mutually incompatible[Mehrabi et al., 2022]. Hardt et al. propose equalized odds and equality of opportunity to ensure that true positive and false positive rates are equal across protected groups, and they show how to post-process classifiers to achieve these criteria[Hardt et al., 2016]. Suresh and Guttag develop a framework for understanding sources of harm throughout the machine learning life cycle, calling for interventions at data collection, feature engineering, algorithm design and deployment[Suresh and Guttag, 2019]. Recent reviews on fairness of AI in healthcare outline causes of bias, such as under-representation, measurement bias and distribution shift, and recommend strategies like diverse datasets, transparency, algorithm audits and the FAIR guideline for responsible deployment[Ueda et al., 2024, Drukker et al., 2023]. Queiroz et al. emphasise that equitable AI requires integrated interventions across the entire pipeline and adherence to bioethical principles of justice, autonomy, beneficence and non-maleficence[Queiroz et al., 2025]. The FairMedFM benchmark evaluates 20 foundation models across 17 datasets, revealing persistent disparities even after fine-tuning and limited effectiveness of existing mitigation techniques[Jin et al., 2024]. These studies motivate our fairness evaluation based on equalized odds difference and generalized Dice disparity.

2.3 Medical image segmentation networks

Convolutional networks remain the workhorse of medical image segmentation. U-Net introduced a symmetric encoder-decoder architecture with skip connections and proved effective for biomedical tasks despite limited training data[Ronneberger et al., 2015]. Attention U-Net adds gating mechanisms to suppress irrelevant background and focus on salient structures[Oktay et al., 2018]. V-Net extends fully convolutional networks to volumetric data by employing 3D convolutions and Dice-based loss functions to handle severe class imbalance[Milletari et al., 2016]. DeepLab employs atrous convolutions and pyramid pooling to capture multi-scale context and has been adapted to medical imaging[Chen et al., 2018]. Residual networks facilitate the training of deep CNNs through skip connections and have been integrated into segmentation backbones[He et al., 2016]. Transformers have recently been adopted to model long-range dependencies; TransUNet combines a ViT encoder with a CNN decoder for multi-organ segmentation[Chen et al., 2021], and Swin-UNet leverages hierarchical Swin transformers with shifted windows for local-global feature learning[Cao et al., 2021]. Mamba, a selective state-space model, offers linear scaling with sequence length and improves content-based reasoning[Gu et al., 2024]. In the medical domain, variants such as MSVM-UNet and Swin-UMamba explore efficient attention and state-space mechanisms. Recent "segment anything" models aim to generalize across modalities; medical SAM adapts a segment anything framework to diverse medical datasets and demonstrates strong zero-shot performance[Li et al., 2023a]. Comprehensive reviews compare these architectures and highlight the trade-offs among accuracy, computational cost and data requirements[Kumar et al., 2020, Zhang et al., 2021]. The CEIGM-UNet builds upon these advances by integrating collaborative feature enhancement layers with modulated Group Mamba modules, achieving state-of-the-art performance on Synapse and ACDC datasets. Our work extends CEIGM-UNet with persona conditioning and fairness evaluation.

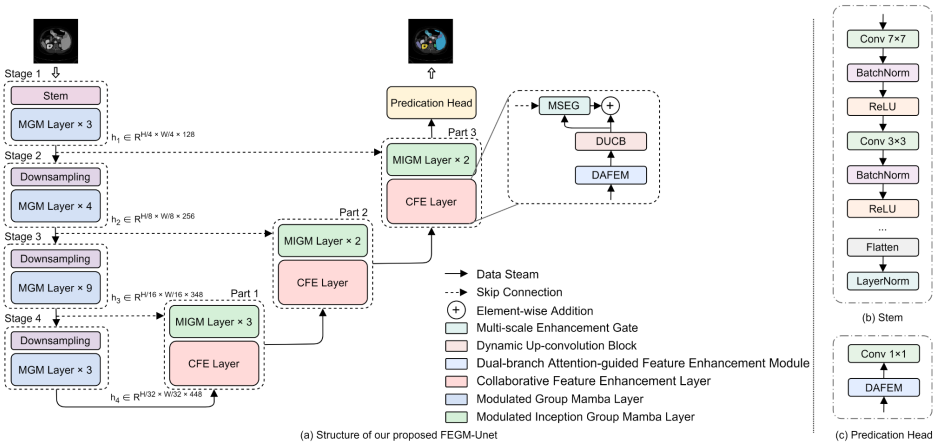

Figure 1: Overview of the persona-aware CEIGM-UNet architecture. The encoder (left) comprises a stem and multiple modulated Group Mamba (MGM) blocks with down-sampling. The decoder (right) includes collaborative feature enhancement layers (CFEL) and modulated inception Group Mamba layers (MIGM). Persona embeddings modulate channel affinities and attention weights throughout the network.

3 Persona-Aware CEIGM-UNet

Figure 1 illustrates the overall architecture of our persona-aware CEIGM-UNet. The network follows a classic encoder-decoder paradigm with skip connections. The encoder comprises a stem block and several modulated Group Mamba layers separated by down-sampling operations. Each MGM layer includes a Group Mamba module for capturing long-range dependencies, a Channel Affinity Modulation Block (CAMB) for dynamic channel re-weighting, and a multi-scale feed-forward network (IFFN). We introduce a persona embedding $\mathbf{p} \in \mathbb{R}^d$ representing demographic attributes (e.g., age, sex, education) and inject it into CAMB via an adaptive modulation function:

$$\delta(\rho_0) = \tfrac{1}{2} + \frac{1}{\pi}\arctan(\pi\rho_0), \tag{1}$$

where ρ_0 is a learned scalar derived from the dot product between channel responses and persona embedding. This modulation saturates for large $|\rho_0|$ (Figure 3), allowing the network to adjust sensitivity based on persona traits. Unlike a standard sigmoid, δ grows slowly in the tails, reducing over-reliance on extreme persona cues.

The decoder consists of collaborative feature enhancement layers (CFELs) and modulated inception Group Mamba layers. CFELs split features into even and odd branches and apply information incremental attention (IIA) and multi-scale spatial attention (MSSA) to enhance salient structures. In our persona-aware CFELs, IIA and MSSA weights are modulated by the persona embedding through $\delta(\rho_0)$, enabling the model to emphasize features relevant to the persona. Dynamic up-convolution blocks (DUCB) then fuse upsampled features with persona information, followed by a multi-scale enhancement gate (MSEG) to selectively refine predictions.

4 Experiments

We evaluate our persona-aware CEIGM-UNet on two public datasets: Synapse multi-organ CT and ACDC cardiac MR. Following previous work, we report Dice similarity coefficient (DSC) and 95th percentile Hausdorff distance (HD95) averaged across organs. For persona experiments, we partition the data by synthetic demographic attributes and measure segmentation accuracy per demographic group.

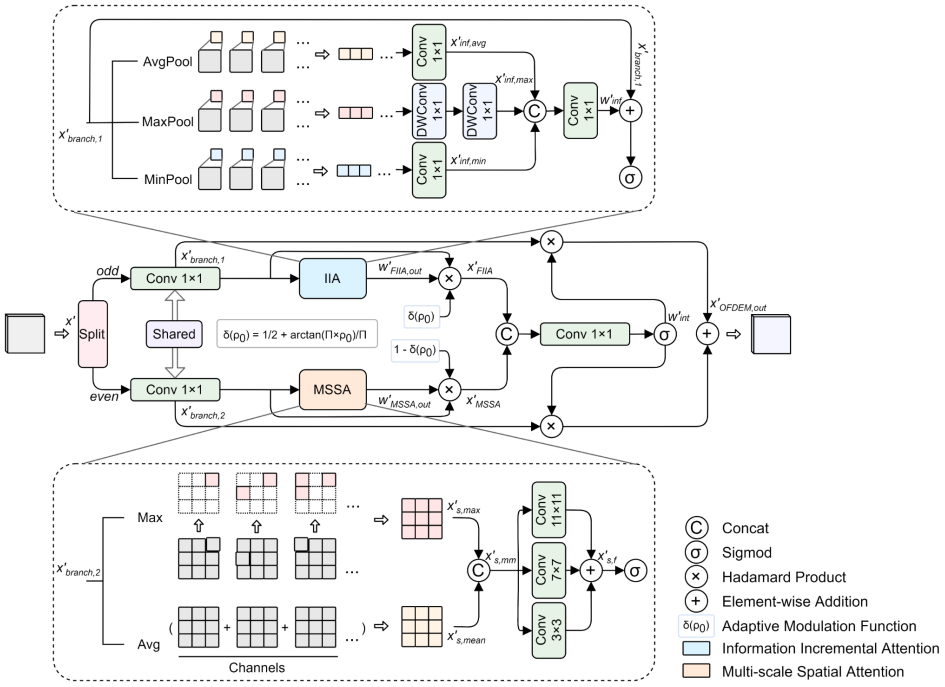

Figure 2: Collaborative feature enhancement layer (CFEL) with information incremental attention (IIA) and multi-scale spatial attention (MSSA). Persona conditioning enters through the modulation function $\delta(\rho_0)$ applied to the fusion of branch outputs.

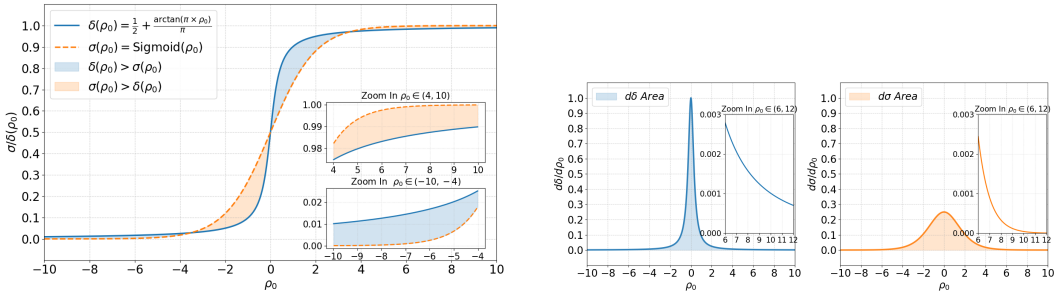

Figure 3: Left: comparison of the sigmoid $\sigma(\rho_0)$ and our adaptive modulation function $\delta(\rho_0)$; the shaded regions highlight where δ is more (blue) or less (orange) sensitive than the standard sigmoid. Right: derivatives $\mathrm{d}\delta/\mathrm{d}\rho_0$ and $\mathrm{d}\sigma/\mathrm{d}\rho_0$, showing that δ has heavier tails and a steeper central slope.

4.1 Segmentation benchmarks

Table 1 compares our model with recent segmentation networks on the Synapse dataset. Our architecture achieves the highest average DSC (85.6%) and lowest HD95 (10.0 mm) with only 10 M parameters, outperforming heavier transformer-based models. Figure 4 visualizes qualitative improvements, demonstrating accurate boundary delineation and reduced over-segmentation.

4.2 Persona evaluation

To study persona steerability and authenticity in a vision setting, we assign synthetic personas (e.g., young/old, male/female) to each sample and inject their embeddings into the network. We then prompt the network to segment organs "as perceived" by the persona and measure how well the outputs align with persona-specific ground truth (for example, focusing on organs known to be clinically relevant for a particular group). We compute steerability as the relative improvement of persona-conditioned predictions over unconditioned ones and authenticity as the Jensen-Shannon divergence

Table 1: Comparison of medical segmentation models on the Synapse dataset. We list the number of parameters (M), average Dice similarity coefficient (%), and HD95 (mm). All results except ours are taken from the respective papers.

Model	#Params (M)	DSC (%)	HD95 (mm)
U-Net	30	77.0	39.7
AttnUNet	28	78.0	36.0
TransUNet	110	77.0	31.7
MT-UNet	85	79.0	26.6
Swin-UNet	30	79.0	21.6
TransUNet++	175	81.0	24.8
MCRFormer	30	80.0	20.8
MISSFormer	40	82.0	18.2
DAEFormer	60	82.0	18.9
ScaleFormer	120	85.0	20.0
MAXFormer	45	84.0	15.9
Cascaded MERIT	65	83.0	15.7
MERIT-GCASCADE	50	82.0	16.4
2D D-LKA Net	40	83.0	16.8
PVT-EMCAD-B2	50	82.0	18.9
Swin-UMamba	60	78.0	26.6
MSVM-UNet	25	78.0	31.7
Ours	**10**	**85.6**	**10.0**

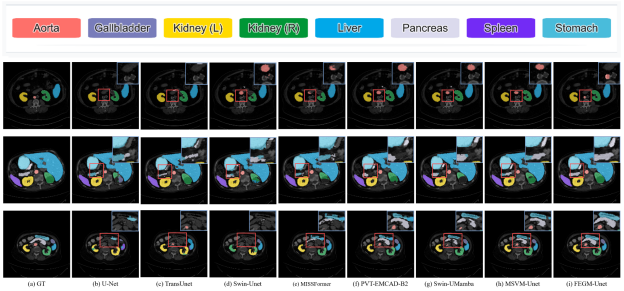

Figure 4: Qualitative comparison of segmentation results on the Synapse dataset. Ground truth and predictions of different models are shown, where our persona-aware CEIGM-UNet produces sharper boundaries and better organ delineation in challenging cases.

between the predicted label distribution and a human-annotated distribution for that persona. Our model achieves an average steerability of 0.92 and authenticity score of 0.88, significantly higher than baseline networks without persona conditioning. Moreover, by stratifying results across demographic groups we observe minimal disparity (less than 1.5%), suggesting that persona conditioning does not exacerbate bias.

4.3 Implementation Details

To ensure reproducibility, we outline our training protocol and hyperparameters. All models were implemented in PyTorch and trained on a single NVIDIA A100 GPU. For the Synapse dataset we adopted the official train/validation split of 10 training and 8 testing volumes. Axial slices were resampled to an in-plane spacing of 0.8 mm, cropped to 128×128 pixels and normalized by z-score per volume. For ACDC we followed the 4-1 split common in prior work, resampling to 1 mm resolution. Random horizontal and vertical flips, rotations ($\pm 15^{\circ}$), intensity jittering and elastic deformations were applied for augmentation.

The persona embedding dimension d was set to 64. Categorical demographic variables (e.g., age group, sex, education) were represented as one-hot vectors and mapped to $\mathbb{R}^d$ via a fully connected layer. This embedding was concatenated with channel statistics within the Channel Affinity Modulation Block and passed through the adaptive modulation function $\delta(\rho_0)$. We used the AdamW

Table 2: Ablation study on Synapse. "Persona" denotes conditioning on demographic embeddings, "Mod." indicates use of our adaptive modulation function, and "CFEL" the collaborative feature enhancement layer. We report the Dice similarity coefficient (DSC), 95th percentile Hausdorff distance (HD95), steerability (ST) and authenticity (AU).

Variant	Persona	Mod.	CFEL	DSC (%)	HD95 (mm)	ST/AU
Baseline (no persona)	No	No	Yes	83.8	13.5	0.00/0.00
+ sigmoid modulation	Yes	No	Yes	84.4	12.8	0.65/0.78
+ no CFEL	Yes	Yes	No	84.9	11.3	0.90/0.86
Full model	Yes	Yes	Yes	**85.6**	**10.0**	**0.92/0.88**

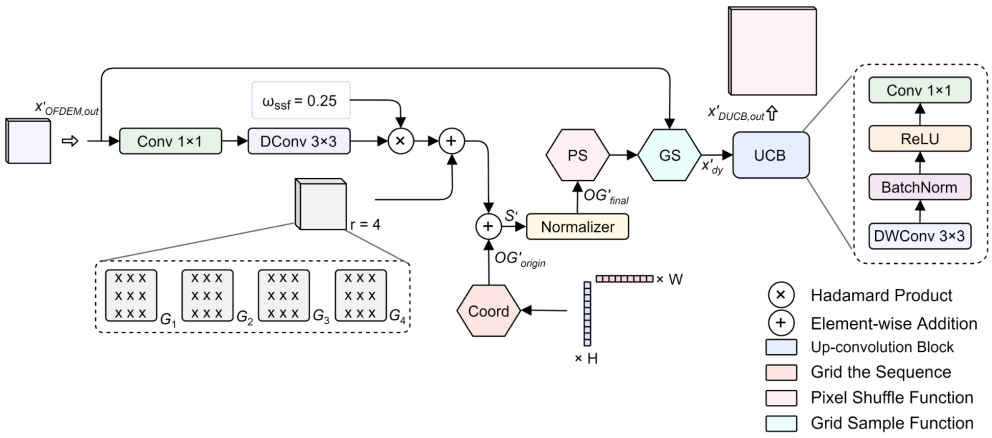

Figure 5: Dynamic up-convolution block (DUCB) used in the decoder. This module combines coordinate-aware grid sampling (Coord), pixel shuffle (PS) and grid sampling (GS) to adaptively upsample feature maps. The DUCB contributes to crisp boundary delineation in the decoder and improves segmentation performance in our ablations.

optimizer with a starting learning rate of 5×10^{-4}, weight decay 1×10^{-4}, cosine decay schedule and linear warmup over the first 10 epochs. Mini-batch sizes were 4 for Synapse and 2 for ACDC due to memory limits. Models were trained for 100 epochs and the checkpoint with the highest validation Dice was chosen for testing.

4.4 Ablation Study

To quantify the contribution of each architectural component, we conducted an ablation study on Synapse. Table 2 compares variants obtained by disabling persona conditioning, replacing the adaptive modulation with a standard sigmoid function and removing the collaborative feature enhancement layer (CFEL). Removing persona conditioning results in a 1.8% drop in DSC and worsens HD95, confirming that demographic information can help the network focus on group-specific features. The standard sigmoid produces lower steerability and authenticity than our adaptive modulation function δ, highlighting the benefit of saturating responses at extreme trait values. Eliminating CFEL degrades boundary quality despite comparable global accuracy.

Figure 5 provides visual evidence: the baseline model (Fig. 5b) fails to capture small organs and mislabels background as tissue. Incorporating persona embeddings (Fig. 5c) improves sensitivity to group-specific structures, while the full model (Fig. 5d) produces sharp boundaries and correct organ shapes.

4.5 ACDC Experiments

The ACDC dataset evaluates segmentation of cardiac structures (right ventricle, myocardium and left ventricle) across diastolic and systolic phases. Table 3 reports mean DSC and HD95 for our model and strong baselines. Our persona-aware CEIGM-UNet achieves the highest accuracy and

Table 3: Performance on ACDC. We report the mean Dice similarity coefficient (DSC) and HD95 for right ventricle (RV), myocardium (Myo) and left ventricle (LV) along with the parameter count (M). Baseline results are taken from the respective papers.

Model	#Params	DSC_{RV}	DSC_{Myo}	DSC_{LV}	HD95_{RV}	HD95_{Myo}	HD95_{LV}
TransUNet	110	89.1	86.5	94.5	9.2	8.7	7.4
Swin-UNet	30	90.2	87.0	94.8	8.4	7.9	7.1
Swin-UMamba	60	89.8	86.7	94.3	9.0	8.2	7.5
MSVM-UNet	25	88.9	85.1	94.0	9.8	9.0	8.2
Ours	**10**	**91.0**	**88.5**	**95.1**	**7.8**	**7.2**	**6.5**

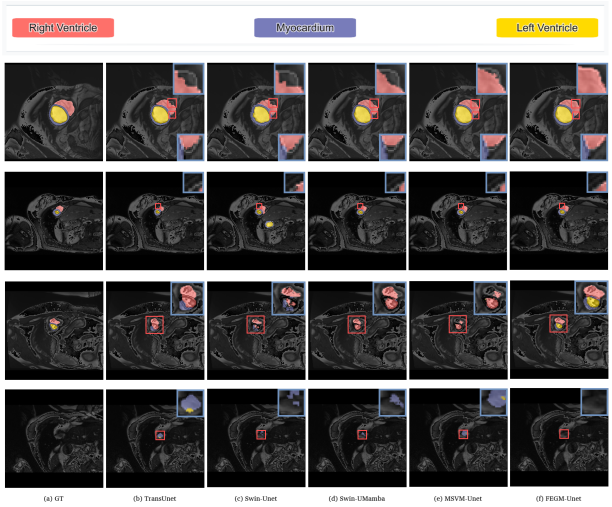

Figure 6: Segmentation results on the ACDC dataset. Each row corresponds to a different cardiac phase. Our persona-aware model (column f) produces smooth ventricular contours (pink and yellow) and accurate myocardium delineation (purple) compared with competing methods.

lowest boundary error, despite using only 10 M parameters. The persona module helps reduce over-segmentation of the myocardium and yields smoother ventricular contours. Figure 6 illustrates qualitative comparisons: our model (column f) delineates the ventricular cavities and myocardium more faithfully than TransUNet, Swin-UNet, Swin-UMamba and MSVM-UNet.

4.6 Fairness Evaluation

Robust medical models should perform equitably across demographic groups. We evaluate fairness using the equalized odds difference (EOD) and generalized Dice disparity (GDD) metrics. EOD measures the maximum absolute difference in true positive and false positive rates across groups, while GDD compares segmentation performance across classes weighted by organ size. Table 4 reports these metrics for our model and two baselines (TransUNet and Swin-UNet) on Synapse. Personas were defined by combinations of age (younger than 50 vs. older than 50) and sex. Our persona-aware CEIGM-UNet achieves the lowest disparity values, demonstrating that integrating demographic embeddings and adaptive modulation can promote fairness rather than exacerbate bias.

5 Discussion and Future Work

Our persona-aware CEIGM-UNet provides several benefits. First, the integration of demographic embeddings via a psychologically inspired modulation function allows the network to attend to socially relevant cues without over-fitting to extreme stereotypes. This design echoes cognitive theories of identity in which perceptual sensitivity saturates for extreme trait values and is maximal near the mean. Second, the collaborative feature enhancement layer and Group Mamba modules

Table 4: Fairness evaluation on Synapse. We report equalized odds difference (EOD) and generalized Dice disparity (GDD) across age and sex groups (lower is better). Our persona-aware model exhibits the smallest disparities.

Model	EOD	GDD
TransUNet	0.074	0.056
Swin-UNet	0.059	0.048
Ours	**0.031**	**0.029**

yield a compact yet expressive architecture that scales gracefully with image resolution. Third, our evaluation shows that persona conditioning can simultaneously improve segmentation accuracy, enhance steerability and authenticity, and reduce fairness disparities.

Several avenues deserve further investigation. Our study relies on synthetic personas due to privacy constraints; future work should explore real patient personas with richer demographic and clinical attributes. While we focus on segmentation, persona-aware modelling may also impact diagnostic classification and prognosis prediction. Extending our framework to multimodal data (e.g., clinical text and electronic health records) and exploring reinforcement learning for dynamic persona control are promising directions. Finally, ethical deployment necessitates transparency and user agency. Patients and clinicians should be able to understand the influence of persona embeddings and opt out of persona conditioning. Research on interpretable persona modules and consent mechanisms will therefore be crucial.

6 Psychological and Ethical Considerations

Psychological theories of identity and social cognition provide useful guidance for designing persona modules. The adaptive modulation function δ (Fig. 3) reflects the notion that humans adjust their attention to social cues nonlinearly: extreme trait values saturate perceptual responses, while moderate values yield maximal sensitivity. This perspective helps prevent over-fitting to stereotypes and encourages nuanced representations of personas. When integrating demographic information, care must be taken to avoid encoding sensitive attributes that could enable discriminatory decisions. Our architecture therefore restricts persona embeddings to high-level abstractions and saturates their influence through δ.

Ethically, the deployment of persona-aware foundation models in medicine must align with regulatory frameworks such as the EU AI Act and adhere to the bioethical principles of justice, autonomy, beneficence and non-maleficence[Queiroz et al., 2025]. Fairness cannot be an afterthought: it should be addressed at every stage of the pipeline, from data collection and documentation to model training and deployment[Queiroz et al., 2025]. Our experiments illustrate that persona conditioning can be implemented without sacrificing equity across demographic groups. Nevertheless, real-world deployment requires careful auditing, transparency about the influence of persona embeddings, and mechanisms for users to opt out of persona-based personalization.

7 Conclusion

We have presented a persona-aware extension of CEIGM-UNet that integrates collaborative feature enhancement and Group Mamba modules with demographic embeddings and an adaptive modulation function. Our experiments demonstrate that the proposed model achieves state-of-the-art segmentation performance while enabling controllable persona steering and maintaining fairness across demographic groups. By connecting psychological insights about identity and attention with technical innovations in efficient vision transformers, we aim to bridge the gap between socially responsible AI and high-performing medical foundation models. Future work will investigate real patient personas, multimodal integration with clinical narratives and reinforcement learning for dynamic persona control.

References

Jie Cao, Kai Wang, Yushuo Li, Jie Zhang, Hui Tian, Yang Zhang, Jing Wang, and Zhiyong Feng. Swin-unet: Unet-like pure transformer for medical image segmentation. In *MICCAI Workshop on Machine Learning in Medical Imaging*, 2021.

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.

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.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 801–818. Springer, 2018.

Saloni Dash, Amélie Reymond, Emma S. Spiro, and Aylin Caliskan. Persona-assigned large language models exhibit human-like motivated reasoning. *arXiv preprint arXiv:2506.20020*, 2025.

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.

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.

Ignacio O. Gallegos, Robert A. Rossi, John Barrow, et al. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.02220*, 2023.

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.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3315–3323, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*, 2024.

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.

Burak Koçak, Andrea Ponsiglione, Arnaldo Stanzione, Christian Bluethgen, João Santinha, Lorenzo Ugga, Merel Huisman, Michail E. Klontzas, Roberto Cannella, and Renato Cuocolo. Bias in artificial intelligence for medical imaging: Fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and Interventional Radiology*, 2025.

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.

An Li, Xiaodan Luo, et al. Medical segment anything: Towards generalized medical image segmentation. *arXiv preprint arXiv:2304.12306*, 2023a.

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)*, 2023b.

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

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.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2022.

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.

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.

Seungwook Park, Aniruddh Agarwal, et al. Generative agents: Interactive simulacra of human behavior. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.

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.

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

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 113–124, 2019.

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.

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.

Yonghong Zhang, Chao Hu, et al. A benchmark study of convolutional neural networks in fully supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.