AEGIS: Anatomy-Embedded Group-Invariant Segmentation for Fair Medical Foundation Models

Sen Wang[1,2], Zhaoyi Zhan[1,2], Zheng Xing[4], Minqing Zhang[5], Chloe Meng Jiang[6], Zhiliang Wang[7], Ruimao Zhang[8], Xiaohui Duan[3]✉, Weibing Zhao[2]✉

[1] Faculty of Computational Mathematics and Cybernetics, Shenzhen MSU-BIT University, Shenzhen, China
[2] Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China
[3] Department of Radiology, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University
[4] Shenzhen University, China [5] The Chinese University of Hong Kong, China
[6] The University of Hong Kong, China [7] Columbia University, United States
[8] School of Electronics and Communication Engineering, Sun Yat-Sen University

Abstract. While foundation models like SAM have advanced medical image segmentation, recent evidence raises safety concerns about subgroup disparities in their medical adaptation. Existing fairness strategies face a critical dilemma when adapting these massive models via Parameter-Efficient Fine-Tuning (PEFT): objective reweighting within PEFT's highly constrained parameter space risks severe accuracy degradation due to feature conflicts, while representation disentanglement often relies on unstable adversarial training that complicates PEFT convergence. To bridge this gap, we propose **AEGIS** (**A**natomy-**E**mbedded **G**roup-**I**nvariant **S**egmentation), a novel framework adapting generalist models for equitable Head and Neck Squamous Cell Carcinoma (HNSCC) segmentation. AEGIS introduces a **Dual-Synergistic** optimization strategy that effectively harmonizes representation purification and dynamic loss adjustment. Specifically, bypassing complex adversarial networks, we introduce an **Attribute-Conditional Feature Alignment** module. Guided by prior attribute metadata, it selectively aligns prototype features: preserving essential biological variations (e.g., gender-specific anatomy) while suppressing non-biological noise (e.g., site-specific scanner variations). Built upon this explicitly disentangled feature space, our **History-Aware Dynamic Weighting** scheme stabilizes reweighting against batch noise via historical temperature scaling, safely targeting long-tail subgroups while maintaining peak generalist performance. Extensive experiments on an in-house multi-center HNSCC cohort and external out-of-distribution (OOD) benchmarks demonstrate that AEGIS achieves a favorable empirical trade-off between segmentation accuracy and group fairness.

Keywords: Medical Foundation Models · Algorithmic Fairness · Parameter-Efficient Fine-Tuning · Head and Neck Squamous Cell Carcinoma.

✉ Corresponding authors: weibingzhao@smbu.edu.cn,duanxh5@mail.sysu.edu.cn
Code available at: https://github.com/lettuce09/AEGIS

1 Introduction

Foundation models built upon the Vision Transformer (ViT) [2] architecture, such as the Segment Anything Model (SAM) [10] and MedSegX [23], have advanced medical image segmentation. However, fairness remains a critical safety concern. Foundation-model-specific evidence shows subgroup disparities in medical segmentation [12], while broader medical-imaging studies document demographic and site-related biases in clinical algorithms [14,15,18,11]. These two lines of evidence suggest that medical foundation models and PEFT pipelines may inherit or amplify existing disparities when adapted to imbalanced clinical cohorts. In Head and Neck Squamous Cell Carcinoma (HNSCC) segmentation [9], such disparities risk inconsistent radiotherapy planning.

Adapting these massive models for specific clinical tasks necessitates Parameter-Efficient Fine-Tuning (PEFT) [6,7,1]. However, achieving fairness under PEFT's strict capacity constraints poses a fundamental dilemma. *Objective reweighting* methods (e.g., GroupDRO [16], FairSeg [20]) force a highly restricted parameter space to reconcile conflicting long-tail distributions, triggering "feature conflicts" that severely degrade generalized anatomical representations. Conversely, *representation disentanglement* methods [4] typically rely on complex adversarial training, leading to highly unstable bi-level optimization and gradient conflicts that are particularly detrimental under PEFT's limited plasticity—a failure mode evidenced even by dedicated PEFT-fairness methods [3].

To bridge this gap, we propose **AEGIS** (**A**natomy-**E**mbedded **G**roup-**I**nvariant **S**egmentation), introducing a **Dual-Synergistic** optimization strategy. Meaningful fairness in foundation model adaptation requires a sequential synergy: one must first purify the representation space, and then dynamically focus optimization on challenging subgroups. AEGIS first utilizes an *Attribute-Conditional Feature Alignment* module to efficiently bypass complex adversarial networks. By selectively aligning prototype features based on attribute types, it explicitly preserves essential biological variations (e.g., gender anatomy) while eliminating non-biological domain noise. Subsequently, a *History-Aware Dynamic Weighting* scheme is applied to this purified feature space, robustly elevating the performance of worst-case subgroups without succumbing to batch-level variance.

Our main contributions are summarized as follows:

(1) We propose AEGIS, a Dual-Synergistic fairness framework for medical foundation models. It breaks the traditional performance-fairness bottleneck in PEFT by coupling representation purification with robust loss reweighting.
(2) We introduce an Attribute-Conditional Feature Alignment module that effectively resolves inter-group feature conflicts, achieving highly efficient feature disentanglement while bypassing the instability of adversarial training.
(3) Extensive experiments on an in-house multi-center HNSCC clinical cohort and external open-world benchmarks demonstrate that AEGIS achieves state-of-the-art results, mitigating bias across gender, age, and site attributes, and placing AEGIS in a favorable high-Dice/low-gap region.

2 Methodology

2.1 Preliminaries: MedSegX and Task Adaptation

Our work builds upon **MedSegX** [23], comprising a frozen Vision Transformer (ViT) [2] backbone initialized from SAM [10] and trainable *Context-aware Mixture-of-Experts (ConMoAE)* adapters. For HNSCC task adaptation via the Parameter-Efficient Fine-Tuning (PEFT) paradigm, we **freeze** the massive ViT image encoder to preserve generalized representations and solely **unfreeze** the ConMoAE adapters and the lightweight mask decoder.

2.2 The AEGIS Framework Overview

Naively fine-tuning foundation models on heterogeneous clinical data leads to fairness degradation. As illustrated in Fig. 1, AEGIS intervenes in the fine-tuning process via a **Dual-Synergistic Optimization** strategy, effectively coupling representation purification and dynamic loss adjustment. At training iteration t, the total loss $\mathcal{L}_{total}^{(t)}$ is defined as:

$$\mathcal{L}_{total}^{(t)} = \mathcal{L}_{task}^{(t)} + \mathbb{1}_{t>T_{warm}} \cdot \left(\lambda_{align}\mathcal{L}_{align}^{(t)} + \lambda_{fair}\mathcal{L}_{fair}^{(t)}\right), \tag{1}$$

where $\mathcal{L}_{task}^{(t)}$ is the standard combination of Dice and Binary Cross-Entropy (BCE) losses acting as the fundamental optimization baseline. The fairness terms $\mathcal{L}_{align}^{(t)}$ and $\mathcal{L}_{fair}^{(t)}$ correspond to our representation alignment and dynamic reweighting modules. The indicator function $\mathbb{1}_{t>T_{warm}}$ activates fairness constraints only after a two-stage warm-up (Sec. 2.5).

2.3 Attribute-Conditional Feature Alignment

A key challenge in fair segmentation is that lesion features often entangle with sensitive attributes. To explicitly disentangle these without GAN instability, we introduce an **Attribute-Conditional Feature Alignment** mechanism over the adapter's feature space. Unconditional global alignment risks collapsing discriminative anatomical structures; thus, while *non-biological style attributes* (e.g., Site/Scanner) should be globally removed, *biological attributes* (e.g., Gender/Age-specific anatomy) must be preserved in the background to maintain structural integrity.

Prototype Extraction via Gating. Let $F \in \mathbb{R}^{C\times H\times W}$ denote the deep feature map extracted from the unfrozen adapters. During training, we use the ground truth mask $Y \in \{0,1\}^{H\times W}$ to spatially pool F into a channel-wise foreground prototype $p_{fg} \in \mathbb{R}^C$ [19,21] and a background prototype $p_{bg} \in \mathbb{R}^C$. To prevent noisy slices from corrupting the representation, we introduce a *Lesion Validity Gating* mechanism:

$$p_{\text{fg}} = \begin{cases} \dfrac{\sum_{h,w} F_{:,h,w}\, Y_{h,w}}{\sum_{h,w} Y_{h,w} + \epsilon}, & \text{if } \sum_{h,w} Y_{h,w} > \tau, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{2}$$

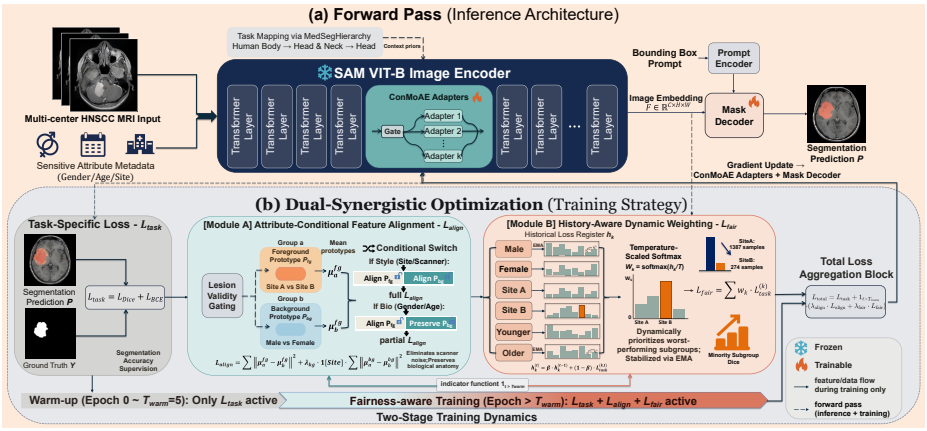

Fig. 1. The overall architecture of the proposed **AEGIS** framework. (a) The forward pass employs a frozen SAM ViT-B Image Encoder with internally trainable ConMoAE Adapters initialized via MedSegHierarchy task mapping, and a trainable Mask Decoder. (b) The **Dual-Synergistic Optimization** strategy harmonizes three objectives: task-specific segmentation loss ($\mathcal{L}_{\text{task}}$); Attribute-Conditional Feature Alignment ($\mathcal{L}_{\text{align}}$), which eliminates scanner noise while preserving biological anatomy via a conditional switch; and History-Aware Dynamic Weighting ($\mathcal{L}_{\text{fair}}$), which prioritizes worst-performing subgroups via EMA-stabilized temperature scaling. Both fairness constraints are activated after a warm-up stage (two-stage training dynamics) to prevent gradient conflicts.

where τ is a pixel threshold (set to 2). The background prototype p_{bg} is computed similarly using $(1 - Y)$.

Conditional Alignment Loss. For subgroups $a, b \in \mathcal{B}$ in the current mini-batch $\mathcal{B}$, we compute group-level mean prototypes μ_a^{fg} and μ_b^{fg} by averaging valid p_{fg}. We formulate a conditional loss that strictly aligns lesion representations while selectively aligning the background:

$$\mathcal{L}_{align} = \sum_{a \neq b} \|\mu_a^{fg} - \mu_b^{fg}\|_2^2 + \lambda_{bg} \cdot \mathbb{1}_{\text{attr} \in \mathcal{S}_{style}} \cdot \sum_{a \neq b} \|\mu_a^{bg} - \mu_b^{bg}\|_2^2, \tag{3}$$

where $\mathcal{S}_{style} = \{\text{Site}\}$. The indicator $\mathbb{1}$ activates background alignment *only* for site-based subgroups. For biological subgroups, this term is safely deactivated to preserve physiological fidelity.

Crucially, our 2D slice-based PEFT paradigm supports large batches (e.g., 64), ensuring intra-batch subgroup diversity and stably computing $\mathcal{L}_{align}$ while avoiding complex cross-batch queues.

2.4 History-Aware Dynamic Weighting

While $\mathcal{L}_{align}$ purifies the representation space, we introduce a **History-Aware Dynamic Weighting** scheme to prevent the model from ignoring harder long-

tail demographics. Assuming the dataset is partitioned into K marginal subgroups, we maintain a **Historical Loss Register** h_k for each subgroup $k \in \{1, \ldots, K\}$, updated using Exponential Moving Average (EMA):

$$h_k^{(t)} = \begin{cases} \beta h_k^{(t-1)} + (1-\beta)\bar{\mathcal{L}}_{task}^{(k,t)}, & \text{if } |S_k| > 0 \text{ in } \mathcal{B} \\ h_k^{(t-1)}, & \text{otherwise} \end{cases} \tag{4}$$

where $\bar{\mathcal{L}}_{task}^{(k,t)}$ is the average task loss for subgroup k in the current batch, and $\beta = 0.9$. All h_k are initialized to 0.

Temperature-Scaled Reweighting. We compute dynamic weights w_k via a softmax function over the historical registers with a temperature T [5]:

$$w_k = \frac{\exp(h_k^{(t)}/T)}{\sum_{j=1}^{K} \exp(h_j^{(t)}/T)}, \quad \mathcal{L}_{fair} = \sum_{k \in \mathcal{B}} w_k \cdot \bar{\mathcal{L}}_{task}^{(k,t)}. \tag{5}$$

By setting T to a low value (e.g., 0.1), AEGIS safely assigns exponentially higher gradients to worst-performing demographics based on historical stability rather than batch-level noise.

2.5 Training Dynamics

To resolve potential gradient conflicts, we employ a **Two-Stage Warm-up Strategy**. In the initial T_{warm} epochs, we purely optimize $\mathcal{L}_{task}$ ($\mathbb{1}_{t>T_{warm}} = 0$). Subsequently, we activate both $\mathcal{L}_{align}$ and $\mathcal{L}_{fair}$ to inject fairness constraints into this matured anatomical feature space.

3 Experiments

3.1 Datasets and Preprocessing

We evaluate AEGIS comprehensively using one in-house and two public out-of-distribution (OOD) datasets. The in-house Head and Neck Squamous Cell Carcinoma (HNSCC) cohort was collected from two collaborative hospitals (2020-2023) with appropriate ethical approval. Representing a multi-center external validation with severe domain shift, this dataset exhibits severe demographic imbalances: Site A contains 1,387 samples (M/F: 1002/385, Age $\leq 60/> 60$: 887/500), whereas the long-tail Site B contains 274 samples (M/F: 226/48, Age $\leq 60/> 60$: 170/104). To further assess fairness generalization, we utilize the Harvard-FairSeg dataset [20] for age and gender shifts, and the PI-CAI multi-national prostate MRI dataset [17] for age and site variations. Following the foundation model's protocol, all 3D MRI volumes were sliced into 2D axial images, resized to 256×256, and normalized. Models predict masks slice by slice, and Dice scores are averaged over samples and then summarized by sensitive subgroup. The HNSCC dataset was randomly split into training and validation sets at an 8:2 ratio.

3.2 Implementation Details and Evaluation Metrics

The framework was implemented in PyTorch and trained on a cluster equipped with NVIDIA RTX 4090 (24GB) GPUs. We employed the MedSegX (ViT-B) architecture, freezing the image encoder while fine-tuning the ConMoAE adapters and mask decoder. Models were optimized using AdamW with a weight decay of 0.01 and an initial learning rate of 5×10^{-5} for 80 epochs. The **Dual-Synergistic** fairness constraints were activated after a 5-epoch warm-up, with loss weights empirically set to $\lambda_{align} = 0.1$, $\lambda_{fair} = 1.0$, and $\lambda_{bg} = 0.05$. Here, λ_{align} balances prototype alignment, λ_{fair} balances history-aware reweighting, and λ_{bg} controls background alignment for style attributes. Crucially, AEGIS is strictly attribute-agnostic during inference, requiring demographic labels only during training. As discussed in Sec. 2.3, our 2D slice-based PEFT paradigm allows us to utilize a large batch size of 64. This naturally ensures diverse subgroup representation within each training iteration, bypassing the need for complex stratified sampling or cross-batch memory queues. For the History-Aware Dynamic Weighting, the lesion-validity threshold was set to $\tau = 2$, the momentum factor to $\beta = 0.9$, and the temperature scaling to $T = 0.1$. We report the Dice Similarity Coefficient (DSC %) for segmentation accuracy and the Equal Accuracy Gap (EA Gap, $\Delta_{\text{EA}} = \max_{g\in\mathcal{G}} \text{Dice}_g - \min_{g\in\mathcal{G}} \text{Dice}_g$) for fairness across Gender (Δ_{Gen}), Age (Δ_{Age}), and Site (Δ_{Site}). EA Gap measures the largest subgroup-level disparity in segmentation accuracy; lower values indicate better group-wise accuracy parity.

3.3 Main Results and OOD Generalization

Table 1 presents the comprehensive quantitative comparison between AEGIS and existing baselines. We interpret overall Dice, subgroup Dice, and EA Gap jointly: a lower EA Gap is favorable only when overall and subgroup-level Dice are preserved rather than uniformly degraded.

Breaking the Performance-Fairness Trade-off: Traditional objective-reweighting methods often suffer from a severe "performance tax" on generalized accuracy when minimizing gaps (e.g., GroupDRO on HNSCC drops overall Dice to 85.9%). This occurs because forcing a restricted parameter space to heavily weight conflicting hard samples disrupts its generalized anatomical representations—precisely the "feature conflict" bottleneck AEGIS is designed to resolve. In contrast, AEGIS explicitly resolves feature conflicts first. As a result, it recovers the generalist performance upper-bound (86.9%) while minimizing bias across all attributes simultaneously, effectively eliminating the Gender Gap ($\leq 0.1\%$) and significantly reducing the Age Gap to 0.9%.

Superior OOD Generalization: In unseen open-world scenarios, AEGIS demonstrates remarkable robustness. On the PI-CAI dataset, where severe domain shifts (e.g., rare scanner protocols) cause massive disparities for baseline models, AEGIS compresses the Site Gap to 1.9% and simultaneously *improves* the overall Dice to 77.8% (surpassing the baseline's 77.4%). Similarly, on the Harvard-FairSeg dataset, AEGIS substantially reduces the Age Gap from 3.1% to 0.9% with negligible impact on overall accuracy.

Table 1. Comprehensive quantitative comparison across multi-center in-domain (HNSCC) and OOD datasets. We report the Overall Dice (%) alongside subgroup-specific Dice and their respective Equal Accuracy Gap (Δ). Best results are in **bold**. '-' denotes metadata unavailable in the dataset. Note: Baseline methods evaluated include nnU-Net [8], MedSAM [13], MedSegX [23], GroupDRO [16], and FlexFair [22]; all fairness-aware methods share the identical MedSegX-PEFT architecture for controlled comparison. External datasets used are Harvard-FairSeg [20] and PI-CAI [17].

Dataset	Method	Overall	Gender (%)			Age (%)			Site (%)		
		Dice↑	Male↑	Fem↑	Δ_G↓	Old↑	Young↑	Δ_A↓	Site A↑	Site B↑	Δ_S↓
HNSCC	nnU-Net	63.4	62.9	63.8	0.9	60.4	65.5	5.1	57.0	66.0	9.0
	MedSAM	77.5	77.2	77.9	0.7	75.8	78.8	3.0	74.3	79.4	5.1
	MedSegX	86.9	**87.0**	86.8	0.2	85.7	**87.6**	1.9	86.9	**87.4**	0.5
	GroupDRO	85.9	85.7	85.9	0.2	85.2	86.6	1.4	85.7	86.2	0.5
	FlexFair	86.5	86.5	86.4	0.1	86.1	87.2	1.1	86.3	86.6	0.3
	AEGIS	**86.9**	86.9	**86.9**	**0.0**	**86.6**	87.5	**0.9**	**87.0**	87.3	**0.3**
Harvard -FairSeg	nnU-Net	79.8	78.5	80.4	1.9	76.5	82.1	5.6	–	–	–
	MedSAM	79.3	80.1	78.9	1.2	76.2	81.5	5.3	–	–	–
	MedSegX	**84.9**	**84.6**	**85.4**	0.8	83.1	**86.2**	3.1	–	–	–
	GroupDRO	83.5	83.4	84.0	0.6	82.5	84.9	2.4	–	–	–
	FlexFair	84.2	83.8	84.4	0.6	83.6	85.4	1.8	–	–	–
	AEGIS	84.6	84.4	84.7	**0.3**	**84.0**	84.9	**0.9**	–	–	–
PI-CAI	nnU-Net	53.9	–	–	–	51.0	55.5	4.5	58.1	48.3	9.8
	MedSAM	73.6	–	–	–	72.1	74.8	2.7	76.0	69.5	6.5
	MedSegX	77.4	–	–	–	76.3	78.2	1.9	79.1	74.5	4.6
	GroupDRO	76.2	–	–	–	75.8	76.6	0.8	78.5	75.2	3.3
	FlexFair	76.9	–	–	–	76.5	77.2	0.7	79.0	76.1	2.9
	AEGIS	**77.8**	–	–	–	**77.7**	**78.2**	**0.5**	**79.3**	**77.4**	**1.9**

Table 2. Ablation study on the HNSCC dataset. **Align**: Attribute-Conditional Feature Alignment; **Weight**: History-Aware Dynamic Weighting.

Align	Weight	Overall Dice ↑	Δ_{Age} ↓	Δ_{Gen} ↓	Δ_{Site} ↓
		86.90	1.90	0.20	0.50
✓		86.55	1.35	0.15	0.36
	✓	86.42	1.48	0.08	0.45
✓	✓	**86.90**	**0.90**	**0.03**	**0.30**

3.4 Ablation Study

To strictly decouple the synergistic contributions of our proposed modules, we conduct an ablation study on the HNSCC dataset (Table 2). Applying solely Attribute-Conditional Alignment improves fairness but drops the overall Dice to 86.55%, as some discriminative capacity is lost during alignment. Conversely, applying only History-Aware Weighting causes a larger accuracy drop (86.42%), confirming that dynamically reweighting an entangled, noisy feature space forces the model to overfit to spurious domain correlations. Integrating both modules yields a sequential synergy: the alignment module first purifies the representation space, providing a robust foundation that enables dynamic weighting to safely focus on empirically underperforming anatomical cases. Consequently, the full AEGIS framework recovers the 86.90% peak accuracy while achieving state-of-the-art fairness.

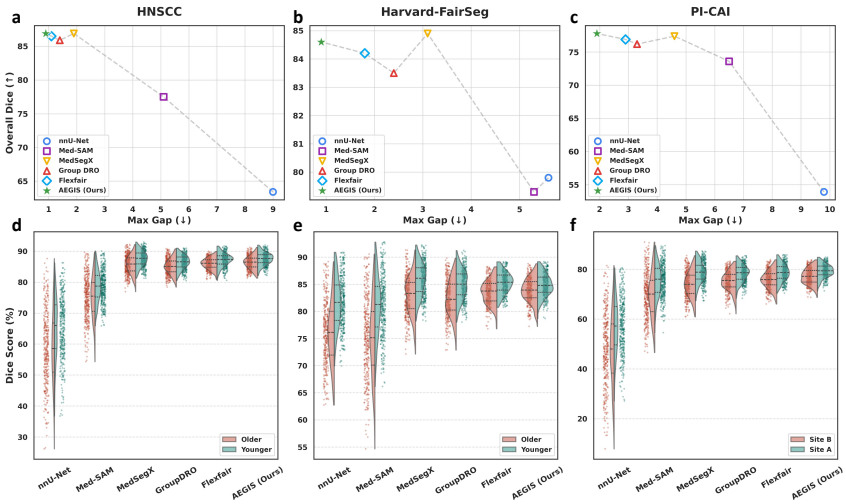

Fig. 2. Comprehensive fairness evaluation across HNSCC, Harvard-FairSeg, and PI-CAI datasets. **(a–c)** Utility-fairness scatter comparisons, where the X-axis denotes the maximum Equal Accuracy Gap across all evaluated sensitive attributes (Gender, Age, and Site), and the Y-axis denotes Overall Dice. Each point represents one trained model under the reported setting; the plot visualizes the empirical utility-fairness trade-off rather than a formal hyperparameter-swept frontier. AEGIS lies in a favorable high-Dice/low-gap region near the ideal top-left corner. **(d–f)** Instance-level Dice distributions (violin plots) for the attribute with the largest performance disparity per dataset: Age subgroups (Older vs. Younger) on HNSCC and Harvard-FairSeg, and Site subgroups (Site A vs. Site B) on PI-CAI. AEGIS significantly elevates the lower-bound performance and brings subgroup distributions closer compared to baseline methods.

3.5 Distributional Stability and Empirical Trade-off

To comprehensively evaluate the macro-level trade-offs and micro-level robustness, we visualize utility-fairness scatter comparisons and instance-level Dice distributions across the three datasets in Fig. 2.

As illustrated in the utility-fairness scatter comparisons (Fig. 2a-c), standard baselines and fairness interventions often reduce subgroup gaps at the cost of overall accuracy. In contrast, AEGIS occupies a highly favorable region towards the top-left corner, substantially mitigating inter-group disparities while maintaining highly competitive SOTA segmentation accuracy across all domains.

Furthermore, the violin plots (Fig. 2d-f) reveal the underlying mechanics of how AEGIS prevents the model from ignoring long-tail groups. Traditional models exhibit severe distributional skewness with abundant "catastrophic failure" cases on hard subgroups. By smoothing batch noise via historical weighting, AEGIS safely elevates the performance lower bound. As shown in Fig. 2(d–f), AEGIS narrows intergroup distributional gaps while substantially truncating the lower tail of minority subgroups.

4 Conclusion

In this work, we address the critical challenge of algorithmic bias in medical foundation models adapted via PEFT. We propose **AEGIS**, a parameter-efficient framework that synergizes representation purification with dynamic loss optimization. By integrating an *Attribute-Conditional Feature Alignment* module and a *History-Aware Dynamic Weighting* scheme, AEGIS explicitly disentangles biological anatomy from spurious scanner noise and stabilizes optimization against batch-level variance, overcoming the traditional performance-fairness bottleneck. Extensive experiments on a multi-center HNSCC cohort and two OOD benchmarks demonstrate that AEGIS effectively mitigates bias across gender, age, and site attributes. The results show a favorable empirical trade-off, elevating the lower-bound performance of long-tail subgroups without compromising generalist segmentation accuracy, providing a robust and scalable pathway for equitable clinical AI deployment.

Acknowledgements This work was partially supported by the Shenzhen Science and Technology Program (Grant No. JCYJ20241202130548062), and the Guangdong Provincial Key Area Projects of General Universities (Grant No. 2024ZDZX1017, 2025ZDZX3049).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.M., Chen, W., et al.: Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. arXiv preprint arXiv:2203.06904 (2022)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
3. Dutt, R., Bohdal, O., Tsaftaris, S.A., Hospedales, T.: Fairtune: Optimizing parameter efficient fine tuning for fairness in medical image analysis. arXiv preprint arXiv:2310.05055 (2023)
4. Gao, Y., Hao, J., Zhou, B.: Fairread: Re-fusing demographic attributes after disentanglement for fair medical image classification. Medical Image Analysis p. 103858 (2025)
5. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
6. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799. PMLR (2019)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)

8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
9. Johnson, D.E., Burtness, B., Leemans, C.R., Lui, V.W.Y., Bauman, J.E., Grandis, J.R.: Head and neck squamous cell carcinoma. Nature reviews Disease primers **6**(1), 92 (2020)
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
11. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020)
12. Li, Q., Zhang, Y., Li, Y., Lyu, J., Liu, M., Sun, L., Sun, M., Li, Q., Mao, W., Wu, X., et al.: An empirical study on the fairness of foundation models for multi-organ image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 432–442. Springer (2024)
13. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
14. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
15. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
16. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=ryxGuJrFvS
17. Saha, A., Twilt, J.J., Bosma, J.S., van Ginneken, B., Yakar, D., Elschot, M., Veltman, J., Fütterer, J., de Rooij, M., Huisman, H.: The pi-cai challenge: public training and development dataset. (No Title) (2022)
18. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature medicine **27**(12), 2176–2182 (2021)
19. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)
20. Tian, Y., Shi, M., Luo, Y., Kouhana, A., Elze, T., Wang, M.: Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. arXiv preprint arXiv:2311.02189 (2023)
21. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 9197–9206 (2019)
22. Xing, H., Sun, R., Ren, J., Wei, J., Feng, C.M., Ding, X., Guo, Z., Wang, Y., Hu, Y., Wei, W., et al.: Achieving flexible fairness metrics in federated medical imaging. Nature Communications **16**(1), 3342 (2025)
23. Zhang, S., Zhang, Q., Zhang, S., Liu, X., Yue, J., Lu, M., Xu, H., Yao, J., Wei, X., Cao, J., et al.: A generalist foundation model and database for open-world medical image segmentation. Nature Biomedical Engineering pp. 1–16 (2025)