Adapting Medical Foundation Models for Coronary Artery Calcium Segmentation from CT Imaging

Jie-En Tsai JOAN040802@GMAIL.COM
Po-Chih Kuo KUOPC@CS.NTHU.EDU.TW
National Tsing Hua University, Hsinchu, Taiwan

Editors: Under Review for MIDL 2026

Abstract

Automated coronary artery calcium (CAC) segmentation plays an important role in coronary artery disease risk stratification; however, conventional task-specific deep learning models typically require large annotated datasets that are challenging to obtain in clinical settings. In this study, we address this limitation by fine-tuning a medical imaging foundation model for CAC segmentation and evaluating its performance under progressively reduced training data. The proposed approach outperforms existing methods when trained on the full dataset and remains competitive even with only 50% of the training data. Moreover, it exhibits smaller performance disparities across sex, age, and CAC severity subgroups compared to a U-Net baseline, highlighting the potential of foundation model fine-tuning for robust and equitable clinical AI applications.

Keywords: Foundation Model, Segmentation, Coronary Artery Calcium, Fairness

1. Introduction

Coronary artery calcium (CAC) is a critical indicator for assessing a patient's risk of coronary artery disease. Accurate measurement of the CAC score is essential for diagnosis and risk stratification (Sharma et al., 2010). Researchers have explored deep learning (DL) approaches to automate CAC scoring (Yee, 2024; Alirr and Khalifa, 2025). Foundation models offer an effective alternative; trained on large-scale and diverse datasets, they can be adapted to downstream tasks (He et al., 2024) with substantially less training data and achieve superior performance to that of task-specific models (Pai et al., 2024).

Beyond prediction performance, ensuring model fairness across patient subgroups is equally important. Studies have shown that AI models trained on datasets with imbalanced demographic distribution tend to exhibit systematic performance disparities against underrepresented patients across demographic subgroups, potentially leading to insufficient care, delayed diagnosis, and violations of bioethical principles (He et al., 2024; Ricci Lara et al., 2022).

In this work, we adapt BiomedParse, a foundation model developed by Microsoft (Zhao et al., 2025), to the CAC segmentation task via transfer learning on a relatively small training dataset. Our objective is to investigate whether a foundation model can match or outperform conventional task-specific models under reduced data requirements, and evaluate model fairness across clinical and demographic subgroups to assess the equitability of the fine-tuned foundation model.

2. Method

Datasets. We use two datasets. The COCA dataset (Stanford University and Stanford AIMI, 2021) contains both gated and non-gated CT images; 437 patients of gated images were used after manual data cleaning. The second dataset is a non-contrast cardiac CT images dataset (Kazemi et al., 2023), which contains 120 subjects (43 patients, 77 healthy people), along with corresponding demographic and CAC-related annotations. We utilized this dataset to analyze model fairness across demographic and clinical subgroups based on the ground truth annotations.

BiomedParse. BiomedParse (Zhao et al., 2025) is a foundation model developed by Microsoft that supports segmentation, detection, and recognition of biomedical objects across nine imaging modalities via text prompts within a single unified framework. We fine-tune BiomedParse to investigate whether the foundation model can achieve performance comparable or superior to that of task-specific models under data-limited conditions.

Phase 1: Low-data fine-tuning. We evaluated performance across four proportions (100%, 75%, 50%, and 25%) using a five-fold cross validation strategy. For each proportion, the held-out test set remained fully intact while only the training folds were downsampled to the target proportion. Each model was trained for 10 epochs, and the final performance was reported as the average across the five folds.

Phase 2: Subgroup fairness evaluation. We used the non-contrast cardiac CT dataset (Kazemi et al., 2023) and evaluated model performance with respect to sex (male, female), CAC score (0, 1–99, 100–299, and ≥300) according to the CAC-DRS classification (Hecht et al., 2018), epicardial tissue volume (quartile-based), and age (0-50, 51-65, >65 years) to reflect different cardiovascular risk stages while maintaining balanced subgroup sizes. To further assess equitability, we adapted the framework (Seyyed-Kalantari et al., 2020), replacing TPR with MAE, RMSE, and bias as performance metrics. For each performance metric M, the performance gap is defined as follows. For binary attributes (i.e., sex), the gap is computed as the difference between the two subgroups: $\text{Gap}_g = M_g - M_{\sim g}$. For non-binary attributes (i.e., CAC score, epicardial tissue volume, age), the gap for each subgroup S_j is defined as its deviation from the median performance across all subgroups: $\text{Gap}_{S_j} = M_{S_j} - \text{Median}(M_{S_1}, \ldots, M_{S_k})$.

3. Results and Discussion

We compared our fine-tuned BiomedParse with a U-Net baseline trained in this study using the same five-fold cross-validation protocol, as well as two Attention U-Net-based approaches (Yee, 2024; Kazemzadeh et al., 2021), and the anatomically guided cascaded U-Net framework with vessel prior (Alirr and Khalifa, 2025), all trained on the COCA dataset (Stanford University and Stanford AIMI, 2021).

When trained on the full dataset, our fine-tuned BiomedParse achieved the highest Dice score among the compared methods, and it slightly outperformed the conventional task-specific approaches and achieved performance comparable to the cascaded framework incorporating vessel priors (Alirr and Khalifa, 2025) with only 50% of the available data.

Regarding the fairness analysis, the fine-tuned BiomedParse generally demonstrates smaller performance gap than the U-Net baseline in sex, age, and CAC score attributes for all three metrics (MAE, RMSE, and bias), suggesting improved equitability across these

clinical subgroups. In contrast, the result for the epicardial tissue volume attribute shows less clear improvement. Both models exhibit subgroup-specific bias across epicardial tissue volume attribute, with the fine-tuned BiomedParse showing more negative bias in the >122 ml group, and the U-Net baseline in the ≤70 ml group.

Table 1: Comparison of Dice scores between the proposed method and prior works for coronary artery calcification segmentation.

Methods	**Dice**
U-Net baseline	0.817
Attention U-Net + Focal Loss (Kazemzadeh et al., 2021)	0.739
Attention U-Net + Gaussian Blur (Yee, 2024)	0.84
Selective U-Net Ensemble (Alirr and Khalifa, 2025)	0.843
Selective U-Net Ensemble with Vessel Prior (Alirr and Khalifa, 2025)	0.855
This work (100% training data)	**0.857**
This work (75% training data)	**0.852**
This work (50% training data)	**0.848**
This work (25% training data)	**0.827**

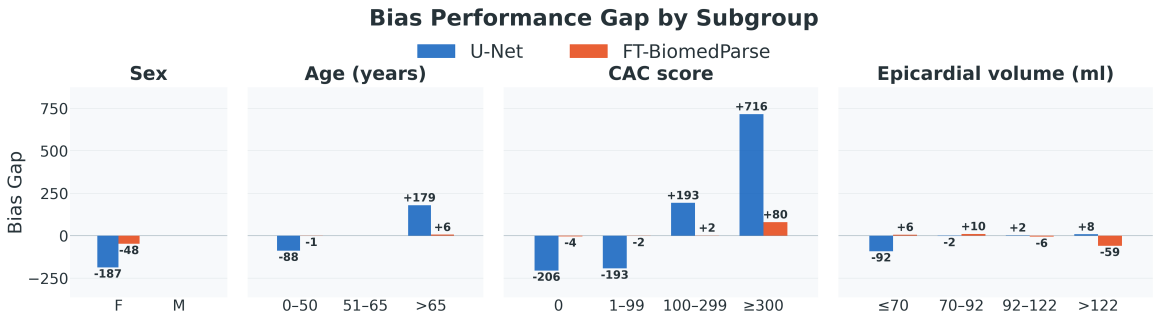

Figure 1: Bias performance gap comparison between U-Net and fine-tuned BiomedParse across subgroups.

4. Conclusion

We fine-tune BiomedParse on the CAC segmentation task and demonstrates that the fine-tuned foundation model achieves performance comparable or superior to conventional task-specific segmentation models while requiring substantially less training data. Notably, the model maintains stable performance even when trained on only 50% of the dataset.

Compared with the U-Net baseline, the fine-tuned model shows reduced performance disparities across most clinical subgroups, suggesting improved equitability. However, systematic bias cannot be fully eliminated, which may influence downstream risk stratification and clinical decision-making. Future work should examine how adjusting training data composition can reduce bias and improve model performance across subgroups.

References

Omar Alirr and Tarek Khalifa. Anatomically guided cascaded u-net ensemble for coronary artery calcification segmentation in cardiac ct. *Bioengineering*, 12(11):1243, 2025.

Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. Foundation model for advancing healthcare: challenges, opportunities and future directions. *IEEE Reviews in Biomedical Engineering*, 18:172–191, 2024.

Harvey S Hecht, Michael J Blaha, Ella A Kazerooni, Ricardo C Cury, Matt Budoff, Jonathon Leipsic, and Leslee Shaw. Cac-drs: coronary artery calcium data and reporting system. an expert consensus document of the society of cardiovascular computed tomography (scct). *Journal of cardiovascular computed tomography*, 12(3):185–191, 2018.

Ali Kazemi, Ahmad Keshtkar, Saeid Rashidi, Naser Aslanabadi, Behrouz Khodadad, and Mahdad Esmaeili. Non-contrast cardiac CT images dataset with coronary artery calcium scoring, 2023. URL https://doi.org/10.17632/msw8kdh348.1.

Sahar Kazemzadeh, Manish Singh, Besher Ashouri, and Samvel Gyurdzhyan. Prediction of coronary artery disease via calcium scoring of chest CTs, 2021. URL https://cs230.stanford.edu/projects_fall_2021/reports/103167584.pdf. CS230 Course Project Report, Stanford University.

Suraj Pai, Dennis Bontempi, Ibrahim Hadzic, Vasco Prudente, Mateo Sokač, Tafadzwa L Chaunzwa, Simon Bernatz, Ahmed Hosny, Raymond H Mak, Nicolai J Birkbak, et al. Foundation model for cancer imaging biomarkers. *Nature machine intelligence*, 6(3): 354–367, 2024.

María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. Addressing fairness in artificial intelligence for medical imaging. *nature communications*, 13(1):4581, 2022.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

Rakesh K Sharma, Rajiv K Sharma, Donald J Voelker, Vibhuti N Singh, Deepak Pahuja, Teresa Nash, and Hanumanth K Reddy. Cardiac risk stratification: role of the coronary calcium score. *Vascular Health and Risk Management*, pages 603–611, 2010.

Stanford University and Stanford AIMI. COCA - coronary calcium and chest CTs. Stanford Center for Artificial Intelligence in Medicine & Imaging, 2021. URL https://doi.org/10.71718/ge5g-ds80.

Nathan Yee. An optimized semantic segmentation deep learning model for automated coronary artery calcification prediction. In *2024 IEEE 5th International Conference on Pattern Recognition and Machine Learning (PRML)*, pages 225–229. IEEE, 2024.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for

joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, 22(1):166–176, 2025.