Investigating the Domain Adaptability of General-Purpose Foundation Models for Left Atrium Segmentation from MR Images

Bipasha Kundu[1(✉)], Bidur Khanal[1], Richard Simon[2], and Cristian A. Linte[1,2]

[1] Center for Imaging Science, RIT, Rochester, NY, USA
bk7944@rit.edu

[2] Biomedical Engineering, RIT, Rochester, NY, USA

Abstract. Segmentation of the left atrium (LA) is crucial for characterizing and appraising left atrial anatomy, morphology, and function in the context of a series of diseases, the most prevalent one being atrial fibrillation (AFib). Despite significant advances in deep learning-based segmentation models, their dependency on large annotated datasets for training limits their effectiveness in niche applications such as atrium segmentation, where annotated data is scarce. Pre-trained foundation models, trained on large-scale general-purpose datasets in a self-supervised manner, can offer an advantage by providing transferable features and enabling adoption to data-scarce domains. In this work, we explore the domain adaptability and robustness of some pre-trained foundation models, such as DINOv2, SAM, and MedSAM, as powerful alternatives for LA segmentation from MRI images. We integrated a modified UNet decoder that leverages the global contextual features encoded by the foundation models. Our approach is evaluated on the 2022 LAScarQS and 2018 LASC segmentation challenge datasets for end-to-end fine-tuning and lower training data settings, respectively. The performance of the UNet decoder was superior to that of the linear decoder used in the original papers of these foundation models, as well as other UNet baselines. Notably, DINOv2 combined with a UNet decoder consistently outperforms the baselines and improves Dice (91.5%, 91.6%) and IoU scores (84.5%, 86.6%), highlighting the model's generalizability and robustness across diverse datasets and limited training data. This study also underscores the transformative potential of foundation models in medical image segmentation, paving the way for more generalized and adaptable solutions across various medical applications.

Keywords: Left atrium segmentation · Vision transformer model · Foundation model · DINOv2 · SAM · MedSAM-v1

1 Introduction

Atrial fibrillation (AFib) is a cardiac condition characterized by irregular heart rhythm that develops in the left atrium (LA), the upper chamber of the heart,

R. Chabiniok et al. (Eds.): FIMH 2025, LNCS 15673, pp. 275–287, 2025.
https://doi.org/10.1007/978-3-031-94562-5_25

significantly increasing the risk of stroke, heart failure, and other heart-related complications [6]. Currently, more than 10 million adult Americans (4.48%) are affected, marking a notable increase from the projections estimated two decades ago [7]. Accurate LA segmentation is essential to identify and quantify scar tissue, which is crucial for effective planning and optimization of pre- and post-treatment, such as ablation therapy. Recent advances in late gadolinium-enhanced magnetic resonance imaging (LGE-MRI) allow for non-invasive visualization of left atrial fibrosis and scar tissue with high resolution located in the LA wall and the pulmonary vein region. LA segmentation is not only critical to understanding the anatomy of the atrium but also serves as a foundational step for subsequent scar segmentation, quantification, and mapping. This work focuses on LA segmentation as a key first step toward a pipeline for left atrial geometry and scar quantification for left atrial ablation therapy planning. As such, the segmentation of the left atrium is our first goal, which is the scope of the work presented here, and will be followed by and integrated with the segmentation, quantification, and mapping of the left atrial scar, which will be reported in a future study, once ready. Therefore, this highlights the importance of precise segmentation techniques to improve diagnostic accuracy and inform effective treatment strategies. While deep learning and ViT-based models have emerged as the leading methods for segmentation tasks [1,14,20,21,32], they are primarily task-specific and require large amounts of annotated data to achieve effective training and prevent overfitting.

Recently, foundation models (FM) have gained significant popularity in natural language processing (NLP) due to their exceptional generalization capabilities. Models like BERT, GPT, and LLaMA are task-agnostic and pre-trained on large datasets to learn general representations, allowing them to perform well across a variety of downstream tasks with minimal fine-tuning [8,23,31].

Following their success in NLP, these models have recently attracted attention in other fields, including computer vision. Popular foundation models like SAM [16], DINOv2 [27] have demonstrated impressive zero-shot performance in natural image tasks and achieved state-of-the-art (SOTA) results when fine-tuned for specific downstream tasks [12,35,37]. In light of their success, these models have recently been employed for various applications in the medical domain. However, medical image segmentation is more challenging due to the delicate anatomical structures, complex object boundaries, and different imaging modalities. Due to the dependency on high-quality prompts per slice, SAM typically underperforms in various medical image segmentation tasks [9,11,24]. Although SAM has generally underperformed in many domains, it prompted multiple efforts to adapt and fine-tune it for the medical domain [10,36]. As an example, MedSAM-v1 [22] was introduced to achieve SOTA performance, but its ability to generalize to other medical imaging modalities remains uncertain. Moreover, the accuracy of the segmentation task is sensitive to the user-defined prompt (size and location of the bounding box and points) of the mask-decoder [25]. Although MedSAM-v2 was introduced in an attempt to further reduce the reliance on per-slice prompts of MedSAM-v1, both architectures exhibit limita-

tions with multiple object scenarios and lead to erroneous segmentation for the continuously changing shape of the left atrium. Lastly, both architectures struggle with the complex boundary regions, size, and contrast of complex anatomy like the left atrium without strong prompt guidance [25,26]. DINOv2, on the other hand, has demonstrated remarkable performance across various medical domains (radiology, X-ray, CT) and organs (kidneys, brain, liver), showing its adaptation, particularly in few-shot settings and diverse adaptation strategies [2,3,33].

Despite promising results, the application of foundation models to LA segmentation remains largely unexplored, especially considering challenges like complex anatomy and the need to identify fibrosis and scar tissue.

This paper focuses on a comprehensive evaluation of recent foundation models while investigating whether a domain gap exists in adopting these pre-trained models for 2D left atrium segmentation. Additionally, designing an effective UNet-like decoder (instead of using a linear decoder) to adapt and decode the rich, task-agnostic features of the foundation model encoder for accurate and efficient segmentation is a key contribution of this paper. Moreover, this work also addresses the challenges of learning with limited labeled data using FMs.

In summary, the contributions of this work are: **(1)** We investigated the domain generalization and effectiveness of encoder of various foundation models (DINOv2, SAM, MedSAM-v1), fine-tuned with two types of decoders; **(2)** We designed a modified UNet decoder on top of each foundation model encoder, leveraging both the shallow and deeper layers of the encoder to capture global and local features effectively; **(3)** We evaluated varying percentages of labeled data on two left atrium datasets, demonstrating that strong decoders ensure precise segmentation and help address the issues of limited annotation.

2 Methods

Our method uses a frozen encoder from pre-trained large vision foundation models (e.g., SAM, MedSAM-v1, DINOv2) with a trainable decoder (linear or UNet-like) trained on the LA segmentation datasets. We aim to assess the transferability of the features learned by these FMs to the specific task of LA segmentation.

2.1 Foundation Models as Encoder

SAM. SAM [16] is an encoder-decoder model by Meta, designed explicitly for promptable segmentation tasks. It was trained on a dataset of 11M high-resolution images and 1 billion associated masks. We chose the SAM-huge variant (636M params). We implemented SAM with the mask decoder and our proposed UNet decoder. While the original resolution of SAM is 1024×1024, we employed positional embedding interpolation to adapt it for a resolution of 448×448, ensuring compatibility with other models. We used forward hooks to extract the intermediate feature maps from the transformer block of the encoder for UNet decoder and a frozen prompt encoder (null prompt) for the mask decoder.

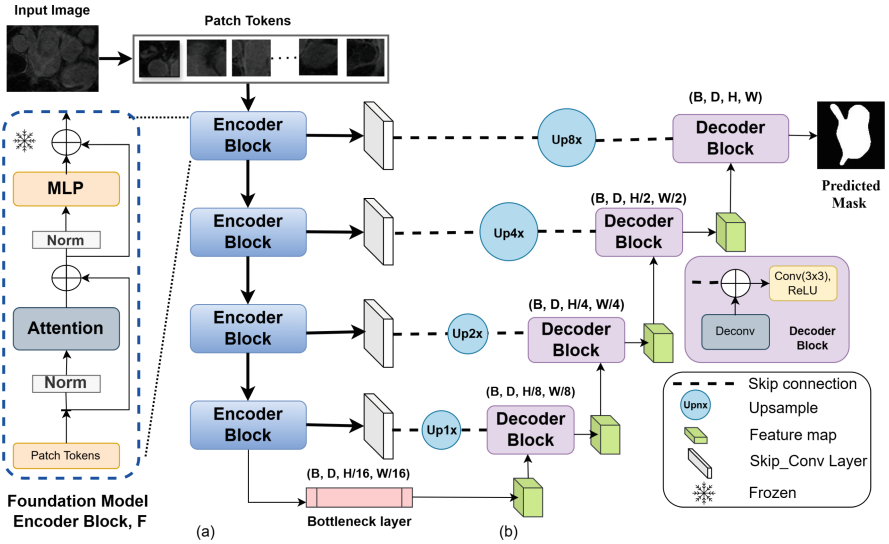

Fig. 1. Overview of the end-to-end fine-tuning framework (a) schematic of transformer layers of the foundational model encoder block; (b) modified-UNet decoder block

MedSAM-V1. MedSAM-v1 [22] is a foundation model with 90.6M parameters for universal medical image segmentation, trained on 1.57M image-mask pairs across 10 modalities. We used the MedSAM-v1 base image encoder and extracted the intermediate features similar to SAM for UNet and the mask decoder.

DINOv2. DINOv2 [27], introduced by Meta, is a state-of-the-art Vision Transformer (ViT)-based model, pre-trained on a diverse dataset of 142M carefully curated natural images. In our work, we experimented with three DINOv2 image encoders: the base model with 88.5M parameters, the large model with 306.4M, and the giant model with 1.1B parameters. Each model is used with its default input resolution to extract robust feature representations.

2.2 Decoder

Linear Decoder. A linear decoder is a simple and lightweight upsampling module that efficiently encodes the aggregated features for prediction with minimal computational overhead. Given an image $x \in \mathbb{R}^{H \times W \times C}$, we split it into 2D patches of size $P \times P$ and flatten each patch into a tokenized 1D vector. $N = \frac{HW}{P^2} + 1$ is the number of tokens, including the [CLS] token. H, W, and C represent the height, width, and channel of the image, respectively. $F_l \in \mathbb{R}^{B \times N \times D}$ represents the features extracted from the encoder, where B is the batch size and D is the embedding dimension. The encoder consists of l layers of Attention and Multi-Layer Perceptron (MLP) blocks. The extracted features are reshaped to $F_l \in \mathbb{R}^{B \times D \times H_l \times W_l}$. The decoder uses four upsampling layers to double the spatial resolution of the reshaped features at each step using bilinear interpolation, followed by 3×3 convolutions, batch normalization, and ReLU activation to yield the final segmentation mask, $\hat{y} \in \mathbb{R}^{B \times 1 \times H \times W}$.

UNet Decoder. A linear decoder captures global context, but struggles to preserve fine-grained spatial details and complex structures, limiting its precision in segmenting irregular anatomical shapes. To address the limitations of the linear decoder in capturing hierarchical spatial information, we designed a UNet-style decoder with the FM encoder. The UNet decoder leverages skip connections to integrate low-level spatial details with high-level encoder representations, thereby preserving fine details and improving spatial consistency. Foundation models excel at capturing global context and high-level patterns, but often lose spatial details due to patch embedding and downsampling operations. Shallow layers capture high-resolution details crucial for boundary refinement, while deeper layers provide abstract, coarse-grained semantic information. The UNet decoder compensates for this loss by using skip connections to merge low-level spatial information from the encoder with high-level abstract features from the foundation model. This combination allows the model to maintain global context and local accuracy. This leads to precise boundary delineation and identification of fine structures, particularly in complex organ segmentation, such as the left atrium.

To take advantage of the complementary strengths of these layers, intermediate feature maps, $F = \{F_1, F_2, \ldots, F_l\}$ are extracted from multiple encoder layers, where $F_l \in \mathbb{R}^{B\times N\times D}$, B is the batch, N is the number of patch tokens, and D is the embedding dimension. Unlike conventional methods, we employ a token-to-feature decoder that excludes the [CLS] token and reshapes patch-level features into $F_l \in \mathbb{R}^{B\times D\times H_l\times W_l}$, where $H_l = \frac{H}{P}$ and $W_l = \frac{W}{P}$, ensuring a focus on relevant spatial information. Instead of mimicking the skip connection like the classic UNet [29], we passed the feature maps through a skip convolution block, comprising a convolutional layer followed by an upsampling operation with scaling factors of n (n = 1x, 2x, 4x, 8x), utilizing a bilinear interpolation algorithm with corner alignments for smoother and accurate spatial resolution doubling. The skip connections are then concatenated with the corresponding decoder blocks. The decoder blocks consist of learnable upsampling blocks and convolutional layers. Each decoder block upsamples the input ($H_l.2^r$,$W_l.2^r$) using bilinear interpolation where r = 1, 2, 3, 4. Then, it concatenates the relevant skip connection F_l and applies 3×3 convolutions followed by batch normalization and ReLU. This process doubles the spatial resolution at each stage, and a final convolutional layer refines the output to generate the segmentation mask. All the FM encoders were frozen during training for both decoders. Unlike traditional approaches that rely solely on the last few layers, our approach incorporates the last layer as a bottleneck. It strategically selects intermediate layers for skip connections to capture hierarchical features ranging from fine-grained spatial details to high-level semantics. For example, in the ViT-base architecture comprising 12 blocks, we use Layer 0, Layer 4, Layer 6, and Layer 8 for skip connections. This multi-level feature fusion approach ensures that global context from deeper layers and local details from shallower layers are effectively combined to facilitate precise segmentation. Figure 1 illustrates the pipeline integrating skip connections, upsampling layers, and convolutional operations to demonstrate its effectiveness.

3 Experiments

3.1 Dataset Description

LAScarQs 2022: The LAScarQs 2022 [17–19] dataset was collected from a challenge hosted by MICCAI. We used Task 2 to evaluate the segmentation of the LA cavity from LGE MRI images. The dataset consists of 130 high-resolution 3D images (576×576 to 640×640 pixels) containing either 44 or 88 slices collected from AFib patients in a clinical setting. Ground truth labels for the LA cavity blood pool were provided.

LASC 2018: The Cardiac Atlas Project provided these data for a MICCAI challenge in 2018 [34] comprising 154 3D GE-MRI scans from patients with AFib. It includes MRI scans and binary segmentation masks (255 = positive, 0 = negative) of the LA cavity, annotated by clinical experts. Each scan has a resolution of $0.625 \times 0.625 \times 0.625\,\text{mm}^3$ and 88 slices along the Z-axis. The dataset is divided into 100 scans with corresponding segmentation masks for training and 54 scans for testing.

For both datasets, we used 10% of the training data as the validation set at the patient level, then extracted 2D slices from the 3D volumes.

3.2 Implementation Details

We used RIT's research computing cluster equipped with NVIDIA A100 GPU to run all the experiments [30]. Before training, all input images were resized to 448×448 for all the methods. We did not apply any pre or post-processing for the FM-based networks and used a batch size of 32 and a learning rate of 0.001 with the Adam optimizer [15] for all FM-based methods with two decoders. We used the original pre- and post-processing methods for SAM/MedSAM-v1 to implement the mask-decoder and an additional post-processing while testing. For the baseline implementation, we followed the same dataset split as the foundation models to maintain consistency. We used a validation set (10% training data) for all our experiments to monitor the overfitting and optimize the model performance. The validation set allowed us to fine-tune hyperparameters and identify the best-performing model checkpoints without biasing the test results. All training setup except nnUNet included a maximum of 50 epochs, and the best validation checkpoints were selected for testing. For nnUNet, we used all the default parameters proposed in [13] with 1000 epochs. We evaluated our experiments using the Dice Similarity Coefficient (DSC) and Jaccard index (IoU). Table 1 summarizes the mean score and standard deviation for both metrics.

3.3 Time Complexity

Table 1 summarizes the training and inference times (in seconds) along with gigaflops (GFLOPs) for the LAScarQS 2022 dataset. Training with a linear decoder was faster than with a UNet decoder because of its simpler architecture. Inference times were calculated for the entire test set used in all methods,

Table 1. Training and Inference Time (seconds) for all Methods on the LAScarQS 2022 Dataset, including both Linear and UNet decoders, along with Inference GFLOPs.

Methods	Training Time (s)		Inference Time (s)		Inference GFLOPs	
	Linear	UNet	Linear	UNet	Linear	UNet
DINOv2-giant	55,540	57,489	246	261	394.0	404.8
DINOv2-large	11,995	13,555	87	89	315.8	325.3
DINOv2-base	6,689	7,526	44	45	92.9	101.8
SAM	29,312	30,029	115	112	495.7	500.4
MedSAM-v1	16,547	17,673	36	37	68.5	83.4
UNet	14,870		106		-	
nnUNet - 2D	59040		323		-	
TransUNet - 2D	5,688		140		-	

with UNet generally requiring slightly more time due to its increased complexity and feature extraction capabilities. The GFLOPs indicate the computational cost of each method, where models with UNet decoders require more operations.

4 Results

Table 2 shows the quantitative comparison of the evaluation metrics for end-to-end fine-tuning, including the baseline models, i.e., UNet [29], nnUNet [13] and Transformer based TransUNet [5]. Foundation models, such as DINOv2, integrated with linear decoders, exhibit minor improvements over baseline methods for the LAScarQS dataset, whereas, for the LASC dataset, nnUNet and TransUNet show better performance than all FM encoder. DINOv2 (all architectures) significantly improves accuracy when paired with our modified UNet decoder. Among all configurations, the DINOv2 models consistently outperform others for both datasets, showcasing their ability to capture fine-grained spatial details.

To ensure a fair and thorough evaluation, we include the performance metrics of both SAM and MedSAM-v1 with their native mask decoders for end-to-end fine-tuning (frozen image and prompt encoder) to highlight the effectiveness of our proposed approach. SAM and MedSAM-v1 perform poorly with their original mask decoders compared to the linear and UNet decoder. The rigid null embeddings of the static prompt encoder retain prior natural image embedding for the center object and sharp edges, whereas the left atrium is an eccentric anatomy with fuzzy boundaries. The mismatch creates incompatible attention patterns in the cross-attention layers ($\text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)$) of the mask decoder, as the frozen prompt embeddings fail to provide domain-appropriate key-value pairs. As a result, end-to-end fine-tuning with a frozen prompt encoder relies solely on adapting the image features, which is insufficient for accurate segmentation, leading to suboptimal performance. This behavior aligns with the original SAM paper that states the effective promptable segmentation requires flexible

Table 2. Quantitative Comparison of DSC (%) and IoU (%) with standard deviation for left atrium segmentation using Linear and modified UNet Decoders,* $p < 0.05$ indicates statistical significance

Decoder Type	Methods	LAScarQS 2022		LASC 2018	
		DSC ↑	IoU ↑	DSC ↑	IoU ↑
Baseline	UNet	84.1 ± 8.3	76.4 ± 12.8	81.9 ± 14.8	74.4 ± 16.5
	nnUNet - 2D	88.7 ± 1.4	83.4 ± 1.5	89.4 ± 1.3	83.6 ± 1.6
	TransUNet - 2D	86.4 ± 3.5	77.1 ± 5.3	87.6 ± 4.13	78.7 ± 6.2
Mask Decoder	SAM-Huge	81.2 ± 22.1	72.2 ± 22.5	82.8 ± 20.0	77.0 ± 20.9
	MedSAM-v1 ViT-base	71.6 ± 27.6	67.4 ± 27.4	75.4 ± 25.5	71.2 ± 26.5
Linear	SAM-Huge	84.4 ± 6.2	75.1 ± 8.4	83.1 ± 13.1	72.8 ± 15.9
	MedSAM-v1 ViT-base	84.2 ± 26.5	74.6 ± 26.1	82.0 ± 21.9	70.5 ± 20.2
	DINOv2 ViT-base	86.9 ± 5.7	80.1 ± 8.4	85.8 ± 12.5	75.8 ± 15.4
	DINOv2 ViT-large	87.0 ± 5.9	80.9 ± 8.5	86.0 ± 13.6	75.8 ± 15.9
	DINOv2 ViT-giant	88.7 ± 5.6	81.1 ± 8.2	86.5 ± 12.4	76.8 ± 15.2
UNet (Ours)	SAM-Huge	86.5 ± 6.4	78.1 ± 9.1	85.1± 14.0	76.0 ± 16.1
	MedSAM-v1 ViT-base	86.3 ± 18.4	77.2 ± 17.8	83.8 ± 19.6	72.7 ± 21.1
	DINOv2 ViT-base	91.3 ± 4.1	83.7 ± 6.9	87.9 ± 12.8	79.0 ± 15.4
	DINOv2 ViT-giant	90.8 ± 5.2	83.2 ± 7.7	88.3 ± 11.8	80.4 ± 14.6
	DINOv2 ViT-large	**91.5 ± 3.9***	**84.5 ± 6.1***	**91.6 ± 10.9***	**86.6 ± 14.1***

adaptation to input prompts, a capability explicitly violated when the prompt encoder is frozen for medical imaging tasks [16].

However, SAM and MedSAM-v1 with a UNet decoder deliver competitive results, closely approaching the performance of the baselines while outperforming other configurations. The use of the UNet decoder significantly boosts segmentation performance compared to linear decoders. Specifically, DINOv2 large with a UNet decoder improves DSC by approximately 2–3% and IoU scores by 1.3–4%. We performed an independent sample Student's t-test between DINOv2 and the best-performing baseline. The results indicate a statistically significant difference ($p < 0.05$) for the LAScarQS 2022 dataset ($p = 0.016$) and the LASC 2018 dataset ($p = 0.003$), suggesting that our method achieves a significantly higher DSC compared to nnUNet. Moreover, the 95% confidence interval (CI) for the mean difference was (0.011, 0.054) for the LASC 2018 dataset and (0.005, 0.052) for the LAScarQS 2022 dataset. Since the confidence intervals do not cross zero, the observed difference is unlikely to be due to random chance.

Furthermore, to evaluate the generalizability of our proposed approach, we applied the trained model of our best-performing method (DINOv2-large) from the LAScarQS 2022 dataset to the LASC 2018 dataset. The model achieved a DSC of 92.6% and an IoU of 87.6% on the LASC 2018 dataset, demonstrating the synergy between UNet's detailed spatial refinement capabilities and their robust feature extraction.

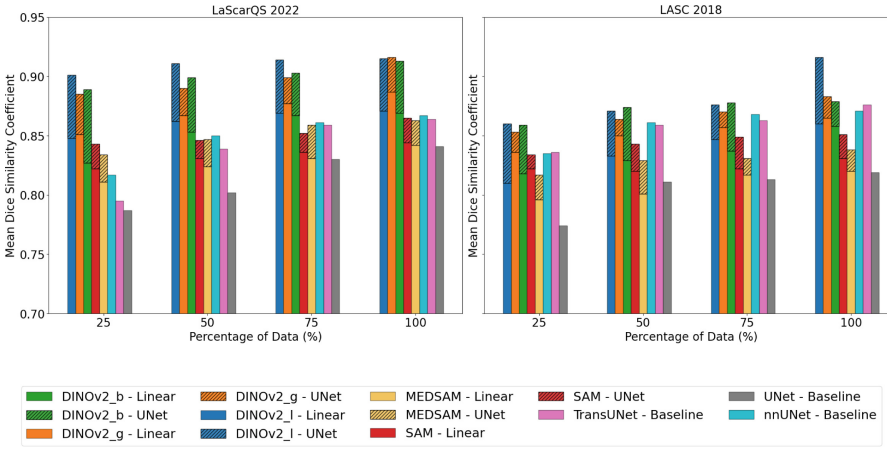

Fig. 2. Comparison of segmentation performance (Mean DSC) across different models and decoder under varying percentages of labeled training data

We also visualize the mean DSC for the varying percentages (25%, 50%, 75%) of training data to show the ability of FMs to perform with limited annotated data in Fig. 2. While the performance appears overall mixed, the results underscore the strength of the FM, particularly DINOv2, SAM, and MedSAM-v1, under limited data conditions. For smaller percentages of training data (25% and 50%), foundation models such as DINOv2 with linear and UNet decoders consistently outperform baseline methods such as nnUNet and TransUNet. Similarly, SAM and MedSAM-v1 with UNet decoders deliver competitive performance in these low-data settings, surpassing the baselines and demonstrating their ability to leverage pre-trained representations effectively. This highlights the robustness of the FM in low-data regimes, where traditional models struggle to generalize effectively. As the percentage of training data increases, the baseline models begin to close the performance gap, particularly in 75% and 100%. However, DINOv2 with the UNet decoder maintains superior performance, showcasing DINOv2's robust global representations. Similarly, SAM and MedSAM-v1, while slightly behind DINOv2, continue to demonstrate strong performance, especially with UNet decoders. In Fig. 3, we present a qualitative comparison of LA segmentation overlays showing true positives (white), false positives (red), and false negatives (blue). The foundation models (DINOv2-large and SAM) trained in 25% of the data are compared to the baseline models (nnUNet and TransUNet) trained in 100%. Among the foundation models, DINOv2-large with a UNet decoder achieves the closest alignment with the ground truth, showing fewer false positives and negatives than SAM and the baselines. Models with linear decoders, such as DINOv2-large and SAM, exhibit higher false positives and less accurate boundary refinements, but their overall performance remains comparable to the baseline. These results underscore the importance of robust decoders

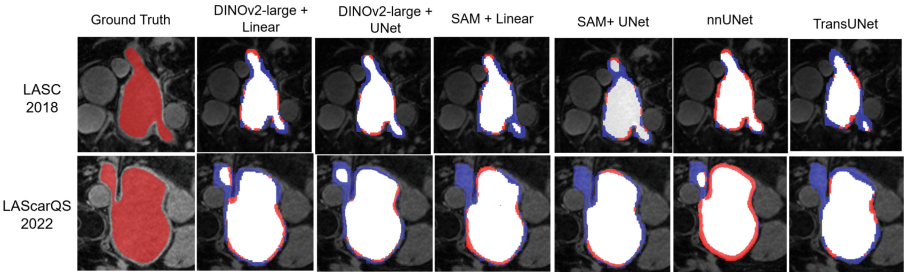

Fig. 3. Qualitative comparison of LA segmentation overlays showing true positive (white), false positive (red), and false negative (blue) regions. Results are shown for DINOv2 and SAM using 25% training data and nnUNet and TransUNet using 100% (Color figure online)

in leveraging foundation models for medical image segmentation, especially in low-data scenarios.

5 Discussion

In this study, we evaluated the domain adaptation of the feature extractor of three foundation models with two decoders, i.e., Linear and UNet, and compared their efficiency with three baselines methods that are considered SOTA techniques for segmentation in the medical domain. The primary motivation for using these two decoders was driven by the rich representations of the characteristics of the encoder of the foundation model. Although the linear decoder captures the global context effectively, the UNet decoder uses the skip connection to reintroduce spatial details and local features from different depths of layers, therefore improving the segmentation performance of the left atrium despite its anatomical challenges.

In comparison with the best-reported results in the original challenges, LASC 2018 achieved the highest DSC of 92.3% [4] and LAScarQS 2022 achieved 89.3% [28] on the validation set. It is important to note that the validation sets and experimental setups of these methods are fundamentally different. The variations in dataset distribution, pre-processing steps, and model training strategies make a direct comparison challenging. Despite these differences, the competitive performance of our method underscores its robustness across different experimental conditions: Our proposed modified decoded method outperforms other baselines on the LAScarQS 2022 dataset and closely matches other baselines on the LASC 2018 dataset.

Our study aims to isolate the contribution of the feature extractor from the segmentation performance by using the same two decoders (Linear and UNet) across all foundation models. This ensures that the segmentation capacity of the model is evaluated consistently in different architectures. The consistent improvement in segmentation outcomes across datasets reflects the adaptability of the encoder-driven feature representation.

However, foundation models like SAM and MedSAM were originally designed with a prompt-based mask decoder that relies on user-defined prompts (bounding boxes, points, and masks) to guide the segmentation task. SAM2 and MedSAM2, being the most recent foundation models, were designed to improve the accuracy and efficiency of their previous versions. Nevertheless, despite being specifically pre-trained on 3D data, they still lead to inconsistent segmentation when generalizing to complex multiobjects without explicit user prompts. As such, our study benchmarks the 2D domain in a prompt-free setting to ensure consistent evaluation across 2D foundation models before extending to the 3D domain.

As part of our future directions, we aim to include different parameter-efficient fine-tuning strategies with additional foundation models to assess their full potential in medical image segmentation.

6 Conclusion

We presented the out-of-the-box potential of natural domain foundation models for LA segmentation in cardiac MRI images. Our primary focus was enhancing domain generalization and robustness while addressing the privacy constraints and ethical challenges of collecting extensive annotated medical data. Our findings show mixed performance, where not all foundation models exhibit outstanding generalization capability in medical domains. DINOv2 consistently outperformed the baselines; SAM and MedSAM-v1 could not compete with the strongest baseline, i.e., nnUNet and TransUNet. However, their performance improved when combined with our modified UNet decoder compared to the linear decoder. Notably, DINOv2 with the UNet decoder can serve as the baseline for transfer learning in medical domain segmentation. The significant improvement in DSC and IoU scores highlights its robustness and reduced dependency on extensive annotated data, demonstrating consistent performance across varying data availability. These findings enable more accurate and efficient segmentation tools for diagnosing and treating complex cardiac conditions like AFib.

Acknowledgments. This work was supported by the National Institutes of Health - National Institute of General Medical Sciences under Award No. R35GM128877 and the National Science Foundation - Division of Chemical, Bioengineering and Transport Systems under Award No. 2245152.

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

References

1. Aghapanah, H., et al.: Cardsegnet: an adaptive hybrid CNN-vision transformer model for heart region segmentation in cardiac MRI. Comput. Med. Imaging Graph. **115**, 102382 (2024)

2. Ayzenberg, L., Giryes, R., Greenspan, H.: Dinov2 based self supervised learning for few shot medical image segmentation. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2024)
3. Baharoon, M., Qureshi, W., Ouyang, J., Xu, Y., Aljouie, A., Peng, W.: Evaluating general purpose vision foundation models for medical image analysis: an experimental study of dinov2 on radiology benchmarks. arXiv preprint arXiv:2312.02366 (2023)
4. Bian, C., et al.: Pyramid network with online hard example mining for accurate left atrium segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart, pp. 237–245. Springer, Cham (2018)
5. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
6. Clinic, M.: Atrial fibrillation symptoms and causes, mayo clinic (2024). https://www.mayoclinic.org/diseases-conditions/atrial-fibrillation/symptoms-causes/syc-20350624. Accessed 15 Dec 2024
7. Clinic, M.: Journal of the American college of cardiology on rhythm disorders & electrophysiology (2024). https://www.jacc.org/doi/abs/10.1016/j.jacc.2024.07.014. Accessed 15 Dec 2024
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers), pp. 4171–4186 (2019)
9. He, S., Bao, R., Li, J., Grant, P., Ou, Y.: Accuracy of segment-anything model (SAM) in medical image segmentation tasks. arXiv preprint arXiv:2304.09324 (2023)
10. He, S., et al.: Computer-vision benchmark segment-anything model (SAM) in medical images: accuracy in 12 datasets. arXiv preprint arXiv:2304.09324 (2023)
11. Huang, Y., et al.: Segment anything model for medical images? Med. Image Anal. **92**, 103061 (2024)
12. Huang, Y., et al.: Comparative analysis of imagenet pre-trained deep learning models and dinov2 in medical imaging classification. In: 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 297–305. IEEE (2024)
13. 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. Nat. Methods **18**(2), 203–211 (2021)
14. Kausar, A., Razzak, I., Shapiai, M.I., Beheshti, A.: 3D shallow deep neural network for fast and precise segmentation of left atrium. Multimed. Syst. **29**(3), 1739–1749 (2023)
15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
17. Li, L., Zimmer, V.A., Schnabel, J.A., Zhuang, X.: AtrialGeneral: domain generalization for left atrial segmentation of multi-center LGE MRIs. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 557–566. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_54
18. Li, L., Zimmer, V.A., Schnabel, J.A., Zhuang, X.: Atrialjsqnet: a new framework for joint segmentation and quantification of left atrium and scars incorporating spatial and shape information. Med. Image Anal. **76**, 102303 (2022)

19. Li, L., Zimmer, V.A., Schnabel, J.A., Zhuang, X.: Medical image analysis on left atrial LGE MRI for atrial fibrillation studies: a review. Med. Image Anal. **77**, 102360 (2022)
20. Lin, H., et al.: Usformer: a small network for left atrium segmentation of 3D LGE MRI. Heliyon **10**(7) (2024)
21. Liu, T., Hou, S., Zhu, J., Zhao, Z., Jiang, H.: Ugformer for robust left atrium and scar segmentation across scanners. In: Challenge on Left Atrial and Scar Quantification and Segmentation, pp. 36–48. Springer, Cham (2022)
22. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nat. Commun. **15**(1), 654 (2024)
23. Mann, B., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165, **1**, 3 (2020)
24. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Med. Image Anal. **89**, 102918 (2023)
25. Mehrnia, M., Elbayumi, M., Elbaz, M.S.: Assessing foundational medical 'segment anything' (med-sam1, med-sam2) deep learning models for left atrial segmentation in 3D LGE MRI. arXiv preprint arXiv:2411.05963 (2024)
26. Mehrnia, M., Elbaz, M.S., et al.: Evaluating foundational 'segment anything' (med-sam1, med-sam2) deep learning models for left atrial segmentation in 3D LGE CMR. J. Cardiovasc. Magn. Reson. **27** (2025)
27. Oquab, M., et al.: Dinov2: learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
28. Punithakumar, K., Noga, M.: Automated segmentation of the left atrium and scar using deep convolutional neural networks. In: Challenge on Left Atrial and Scar Quantification and Segmentation, pp. 145–152. Springer, Cham (2022)
29. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
30. R.I. of Technology: Research computing services (2019). https://doi.org/10.34788/0S3G-QD15. https://www.rit.edu/researchcomputing/
31. Touvron, H., et al.: Llama: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
32. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: a survey. IET Image Proc. **16**(5), 1243–1267 (2022)
33. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. arXiv preprint arXiv:2308.02463 (2023)
34. Xiong, Z., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Med. Image Anal. **67**, 101832 (2021)
35. Zhang, C., et al.: A survey on segment anything model (SAM): vision foundation model meets prompt engineering. arXiv preprint arXiv:2306.06211 (2023)
36. Zhang, Y., Shen, Z., Jiao, R.: Segment anything model for medical image segmentation: current applications and future directions. Comput. Biol. Med. 108238 (2024)
37. Zhou, T., et al.: Image segmentation in foundation model era: a survey. arXiv preprint arXiv:2408.12957 (2024)