Fourier Contour Learning for Efficient and Traceable Cardiac MRI Quantification

Yi Yu[1*], **Zhenyu Bu**[1*], **Ziyu Zhang**[2], **Yixuan Liu**[1], **Aotian Chen**[1], **James Mihalich**[1], **Parker Martin**[1], **Xia Ning**[1], **Yuchi Han**[1], **Yuan Xue**[1†]
[1]The Ohio State University [2]Nanjing University
xue.643@osu.edu
https://github.com/yuyi1005/cardiocontour

Abstract

Automated cardiac MRI quantification is commonly framed either as dense segmentation followed by voxel counting or as direct regression of global measurements. The two paradigms are difficult to unify: segmentation is visually interpretable but label-intensive and output-heavy, while regression is compact but less verifiable. We propose Fourier Contour Learning, which represents cardiac boundaries with compact anatomical Fourier contours. The predicted coefficients define smooth, interpretable contours and enable differentiable volume computation, connecting mask supervision, measurement supervision, and unlabeled-image regularization within one framework. We instantiate this representation with an anatomy-query decoder and cross-slice attention which creates a traceable path from image to contour to clinical measurement. Across three public datasets and an in-house all-phase cohort, our method achieves segmentation accuracy comparable to strong dense-prediction baselines, improves boundary agreement, and reduces clinically relevant volume error while using only a small number of output parameters per structure. We further show that the learned model provides an efficient interface for vision-language question answering and surface reconstruction. These results demonstrate the potential of Fourier contours as a unified intermediate representation for scalable, traceable, and clinically grounded cardiac analysis.

1 Introduction

Cardiac magnetic resonance (CMR) is the clinical reference standard for quantifying cardiac function, with ventricular volumes such as end-diastolic volume (EDV), end-systolic volume (ESV), and ejection fraction (EF) serving as key biomarkers for diagnosis and prognosis. A large body of automated methods has been developed to replace manual workflows, substantially improving efficiency and reproducibility [1, 2, 3]. Recent advances in computer vision and deep learning further enable models to learn spatial and temporal representations directly from CMR data, supporting more accurate and scalable cardiac quantification [4].

Existing approaches can be broadly categorized into segmentation-based [5, 6] and regression-based [7] methods. Segmentation-based methods first predict pixel-wise masks of cardiac structures and then compute volumes by integrating positive voxels. In contrast, regression-based methods directly predict global metrics from images in an end-to-end manner.

While both paradigms have achieved notable progress, they exhibit inherent limitations, as shown in Fig. 1. Segmentation methods rely on supervision from pixel-wise annotations, which are costly and scarce. Moreover, a single 256×256 mask contains over 65K spatial elements, and the output

[*]Equal contribution.
[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Methods	Segmentation-based			Regression-based			Contour-based (ours)		
Visual Explainability			Mask		*LV: 90 mL RV: 60 mL ...*	N/A			Contour
Supervision Flexibility	Mask	Measurement	Unlabeled	Mask	Measurement	Unlabeled	Mask	Measurement	Unlabeled
Representation Efficiency	256 Dense Tokens		65.5k Parameters	1 Global Token		3 Parameters	3 Anatomical Tokens		3×12 Parameters
Anatomical Continuity	*Implicit (learned)*			*Weak (not enforced)*			*Explicit (continuous region)*		
Differentiable Volume	*Via mask voxel counting*			*Direct volume regression*			*Analytic from coefficients*		
Accuracy (ACDC – LV)	93.43 Dice (%)	90.97 NSD (%)	5.76 ΔVol (mL)	N/A Dice (%)	N/A NSD (%)	13.25 ΔVol (mL)	93.65 Dice (%)	94.78 NSD (%)	5.24 ΔVol (mL)

Figure 1: Output representations for CMR quantification. Segmentation predicts dense masks that are visually verifiable but output-heavy; regression predicts compact measurements but lacks explicit anatomy; Fourier contour learning predicts compact anatomical contours whose coefficients decode to boundaries and analytically determine volumes. nnU-Net [15] and Germain et al. [7] are shown as representative segmentation and regression methods. Efficiency is evaluated per 256×256 slices.

size grows rapidly for volumetric or cine data. Regression methods are more compact, but they do not explicitly model anatomy. As a result, their predictions are harder to audit visually and are only weakly constrained by spatial evidence.

These limitations are fundamentally tied to two key challenges: **1) Data availability.** Pixel-wise annotations are scarce. Taking ACDC [1] as an example, only 200 out of 2,898 3D volumes are annotated with segmentation masks, restricted to ED and ES frames. In contrast, UK Biobank [8] includes over 96,000 subjects providing primarily volumetric measurements. Standard supervised segmentation cannot directly exploit such measurement-labeled or unlabeled data. **2) Token efficiency.** Dense anatomical outputs are inefficient for emerging vision-language model (VLM) workflows [9, 10, 11, 12]. When integrating anatomical information into a medical VLM, a $256 \times 256 \times 10$ mask corresponds to 2.6K tokens, with a full cine of 30 frames resulting in over 76K tokens. It significantly increases computational cost and limits scalability, raising a central question: *Can we design a representation that **1)** bridges segmentation and regression, **2)** remains traceable while being compact, and **3)** is compatible with heterogeneous supervision?*

We answer this question with Fourier contour learning. Instead of predicting dense masks or scalar measurements, our model predicts a small set of Fourier contour coefficients for cardiac anatomy (see Fig. 2). These coefficients decode to smooth, visually inspectable contours and analytically determine slice areas and ventricular volumes. In line with 3D representation learning [13, 14], our contour tokens also provide a compact representation of 3D cardiac anatomy, with potential applications in surface construction and downstream vision-language reasoning (see Apps. C and D).

Our main contributions are summarized as follows: **1)** We introduce a Fourier contour representation for CMR quantification that bridges dense segmentation and direct regression through compact, traceable anatomical coefficients. **2)** We develop a simple yet effective query-based contour prediction framework with analytic measurement supervision and self-supervised contour regularization, enabling unified learning from mask-labeled, measurement-labeled, and unlabeled CMR data. **3)** Extensive experiments are carried out to analyze design choices and evaluate performance on three public datasets and an in-house all-phase cohort.

2 Related Work

Segmentation-Based CMR Analysis. As a fundamental task in computer vision, segmentation has been extensively studied and serves as the basis for many downstream medical imaging tasks. Convolutional architectures, ranging from U-Net [16] to nnU-Net [15], established strong performance. To better capture long-range anatomical dependencies, Transformer-based and hybrid architectures were introduced [17, 18, 19, 20, 21, 22]. More recently, Mamba-based models further explored global context modeling [23]. Segmentation is widely adopted for CMR quantification, where volumes and cardiac functions can be derived from segmented structures by counting foreground voxels [6, 24].

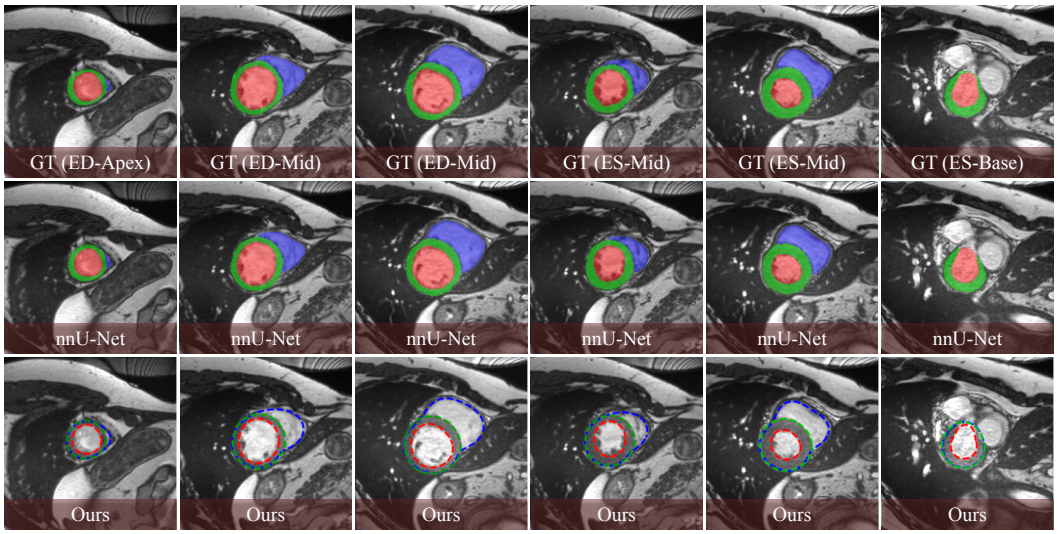

Figure 2: Qualitative comparison with dense segmentation. On representative ACDC slices from basal, mid-ventricular, and apical levels at ED and ES, Fourier Contour Learning produces smooth and coherent anatomical boundaries with visual quality comparable to nnU-Net [15], while using a compact contour parameterization. Unlike dense masks, the predicted contours provide ordered boundaries that can be directly inspected and analytically converted into volume measurements.

Despite the success, several limitations are observed: **1)** They rely heavily on pixel-wise annotations, limiting scalability. **2)** Dense masks introduce substantial token redundancy in VLM settings.

Regression-Based CMR Quantification. To reduce the dependence on dense annotations, some studies have explored direct, segmentation-free regression of ventricular volumes or ejection fraction from CMR. Germain et al. [7] employed a 3D ResNet-18 [25] to regress cardiac volumes from image data. Some approaches incorporated temporal modeling or dynamically constructed networks to improve estimation [26, 27, 28, 29]. These methods simplify the measurement pipeline by avoiding dense segmentation, but also introduce several limitations: **1)** They lack spatial and visual traceability, as no explicit anatomical structures are modeled. **2)** Regression relies on sparse global supervision signals, leading to weak spatial constraints, slower convergence, and reduced robustness.

Multi-Task Foundational CMR Models. It is another active research direction, featured by leveraging autoencoder to pretrain backbones with large-scale unlabeled data. For example, CineMA [4] is pretrained on 74,916 cine CMR studies and subsequently fine-tuned for 23 downstream tasks, including segmentation and volume regression. Similar directions appear in [30, 31, 32, 4]. These works demonstrate the effectiveness of utilizing unlabeled data. However, although such frameworks support multiple tasks, segmentation and volume estimation are essentially optimized independently, lacking explicit anatomical correspondence. As a result, these methods may achieve visually accurate segmentation while producing inaccurate volumes. In contrast, our method serves as an explicit bridge between segmentation and regression through a unified anatomical representation.

Fourier Contour Learning. Fourier descriptors provide compact parametric representations for closed shapes. Recent learning-based methods have revisited this idea with neural networks. For example, [33] predicts contours of remote sensing objects, while CPN [34] and EFDTR [35] incorporate Fourier representations to improve segmentation performance. In medical imaging, Fourier contours have also been explored for enhancing shape consistency and domain generalization [36, 37, 23]. After reviewing these studies, we find that contour representations are well suited for cardiac structures, while their application in the CMR domain remains largely unexplored, motivating this work.

3 Methods

We present Fourier Contour Learning (see Fig. 3), a framework that changes the representation of CMR quantification from dense masks or scalar measurements to compact anatomical contours. For each slice, the model predicts a small set of Fourier coefficients that can be decoded into visually

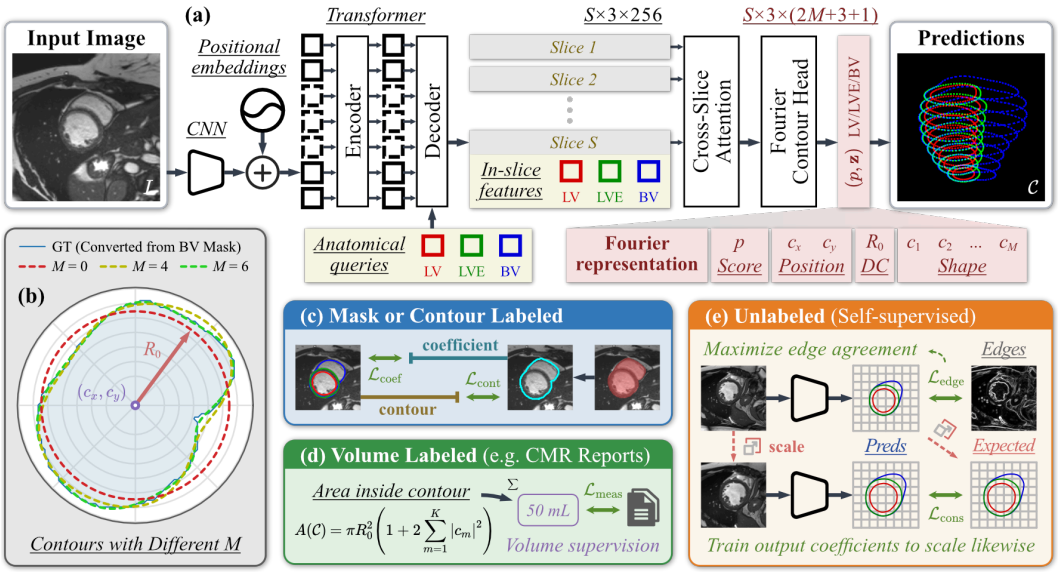

Figure 3: Method overview. **(a)** DETR-style architecture used to predict Fourier contour representations, with cross-slice attention to aggregate 3D context. **(b)** Contours with different truncation orders M. **(c–e)** Training with heterogeneous supervision. **(c)** Mask-labeled data supervising contours and coefficients. **(d)** Volume-labeled data supervising via the enclosed area equation. **(e)** Unlabeled data utilized through self-supervised learning, including edge alignment and cross-scale consistency.

inspectable boundaries and analytically converted into ventricular volumes. Our framework supports learning from dense masks, scalar measurements, and unlabeled images within a single model.

3.1 Fourier Representation for Cardiac Contours

We predict three closed contours: the left ventricular endocardial contour (LV), the left ventricular epicardial contour (LVE), and the biventricular outer contour (BV) enclosing the LV epicardium and right ventricle (RV). The myocardium (MYO) is defined as the region between LV and LVE, and the RV is obtained from the region between LVE and BV. Given a reference center (c_x, c_y), contour $\mathcal{C} = \{(x, y)\}$ can be expressed in polar coordinates by a radius function $r(\theta)$, where $\theta \in [0, 2\pi]$ denotes the angle and $r(\theta)$ is the distance from (c_x, c_y) to the contour along direction θ. Specifically, the contour points can be written as:

$$x(\theta) = c_x + r(\theta)\cos\theta, \quad y(\theta) = c_y + r(\theta)\sin\theta, \tag{1}$$

where $r(\theta)$ is periodic and can be represented by a truncated complex Fourier series, with $R_m \in \mathbb{C}$ denoting the Fourier coefficient at frequency m, and M the number of retained frequencies:

$$r(\theta) = \sum_{m=-M}^{M} R_m e^{im\theta}. \tag{2}$$

Contour-to-Coefficient Encoding. Given a contour $\mathcal{C}$, to obtain its Fourier representation $\mathbf{z}$, we first compute its reference center from the bounding extent: $c_x = (\min(x) + \max(x))/2$, $c_y = (\min(y) + \max(y))/2$. Then the contour is converted into a polar function $r(\theta)$ by sampling radial distances from (c_x, c_y) at uniformly spaced angles. The Fourier coefficients are computed via fast Fourier transform (FFT). For real-valued radius functions, the coefficients satisfy conjugate symmetry $R_{-m} = \overline{R_m}$. Therefore, we parameterize the contour using the following coefficient vector:

$$\mathbf{z} = (c_x,\, c_y,\, R_0,\, \mathrm{Re}(c_1),\, \mathrm{Im}(c_1),\, \ldots,\, \mathrm{Re}(c_M),\, \mathrm{Im}(c_M))\,, \tag{3}$$

where $R_0 \in \mathbb{R}$ captures the average radius of contour, while the normalized coefficients $c_m = R_m/R_0$ encode scale-invariant shape geometry. Under this formulation, contour $\mathcal{C}$ is compactly represented by a finite set of Fourier coefficients, which enables efficient encoding, decoding, and direct geometric computations. The decoding from $\mathbf{z}$ to contour $\mathcal{C} = \{(x, y)\}$ are defined by Eqs. (1) and (2).

Area Inside Contour. Given the radial Fourier representation, the enclosed area of the contour $\mathcal{C}$ can be computed directly from its coefficients without explicitly reconstructing the boundary:

$$A(\mathcal{C}) = \frac{1}{2}\int_0^{2\pi} r(\theta)^2 d\theta = \pi \sum_{m=-M}^{M} |R_m|^2 = \pi R_0^2 \left(1 + 2\sum_{m=1}^{M} |c_m|^2\right), \tag{4}$$

which can be computed analytically from the predicted parameters $\mathbf{z}$, enabling differentiable measurement supervision. Cardiac volumes are then obtained by summing across slices.

3.2 Anatomy-Query Contour Prediction

We adopt an encoder–decoder architecture inspired by DETR [38], as shown in Fig. 3(a). During preprocessing, images are kept at their original resolution when the long edge is no larger than 256 and zero-padded to 256×256 for batch processing. If the long edge exceeds 256, it is resized to 256 while preserving the aspect ratio, followed by padding. The image intensities are normalized to $[0, 1]$. Spatial coordinates are normalized to $[0, 1]$ along both axes as in DETR.

The input image, denoted by I, is first processed by a convolutional neural network (CNN) feature extractor, followed by the addition of positional embeddings. The resulting features are then fed into a Transformer encoder-decoder. Unlike standard DETR with 100 object queries for detection, we use three randomly initialized anatomical queries corresponding to LV, LVE, and BV. Each query predicts one contour, so no bipartite matching is required.

Cross-Slice Attention. To support both 2D and 3D inputs in a unified manner, we introduce a cross-slice attention. Given a stack of S slices, the decoder queries produce in-slice features $\mathbf{X} \in \mathbb{R}^{S\times Q\times C}$, where $Q = 3$ denotes the number of queries and $C = 256$ is the feature dimension. The lightweight cross-slice attention module exchanges information along the slice dimension, allowing the model to use 3D context while respecting the anisotropic spacing of CMR slice stacks. This formulation naturally reduces to 2D when $S = 1$, while enabling volumetric reasoning when multiple slices are available. It is particularly suited to CMR data, which are typically acquired as slice stacks with large inter-slice spacing rather than isotropic 3D volumes. Details can be found in App. A.

Fourier Contour Head. The features output by cross-slice attention, $\mathbf{X}' \in \mathbb{R}^{S\times Q\times C}$, are then fed into the Fourier contour head to predict contour-presence score $p \in \mathbb{R}^{S\times Q\times 1}$ and coefficients $\mathbf{Z} \in \mathbb{R}^{S\times Q\times(2M+3)}$. This corresponds to $S \times Q$ contour parameter vectors, denoted as $\mathbf{z}^{q,s} \in \mathbb{R}^{2M+3}$ for each slice s and query $q \in \{\mathrm{LV}, \mathrm{LVE}, \mathrm{BV}\}$. Details can be found in App. A.

Default Model Settings. For the convolutional backbone model, we choose ResNet-50 [39] unless ConvNeXt [40] is specified. For modules requiring contour sampling, the number of sampled points is set to $K = 64$, the Fourier representation uses $M = 4$ frequencies unless otherwise specified.

3.3 Self-Supervised Contour Regularization: Image–Contour Edge Alignment

Image edges have been used for unsupervised boundary regularization [41]. Similarly, we introduce an edge alignment loss that aligns predicted contour normals with image gradients to encourage anatomically plausible boundaries without pixel-wise labels (see Fig. 4). We first compute the gradient field $\mathbf{G} = (G_x, G_y)$ of the input image using the Sobel operator, where G_x and G_y represent the horizontal and vertical image gradients, respectively. Given the predicted contour parameterized by $\mathbf{z}$, we reconstruct a discrete contour by uniformly sampling K points along θ, denoted as $\{\mathbf{q}_k\}_{k=1}^{K}$. At each sampled point $\mathbf{q}_k$, we compute the unit normal vector $\mathbf{n}_k \in \mathbb{R}^2$.

Polarity of Contour Normals. We assign the normal direction based on expected intensity transitions. For LV and BV, where the blood pool is brighter than surrounding tissue, we use inward-pointing normals. For LVE, which appears darker, we use outward-pointing normals. This polarity-aware design ensures consistent gradient alignment and prevents mismatches between adjacent structures (e.g., LV and LVE), as illustrated by the clockwise/counter-clockwise orientations in Fig. 4.

Contour Normal Field. These discrete normals are then rasterized into a dense contour normal field $\mathbf{F} = (F_x, F_y)$, defined over the image grid. Specifically, for each pixel location (i, j), the normal field is constructed by aggregating the contributions from all sampled contour points:

$$\mathbf{F}_{ij} = \sum_{k=1}^{K} \exp\left(-\frac{\|[i,j] - \mathbf{q}_k\|^2}{2\sigma^2}\right)\mathbf{n}_k, \tag{5}$$

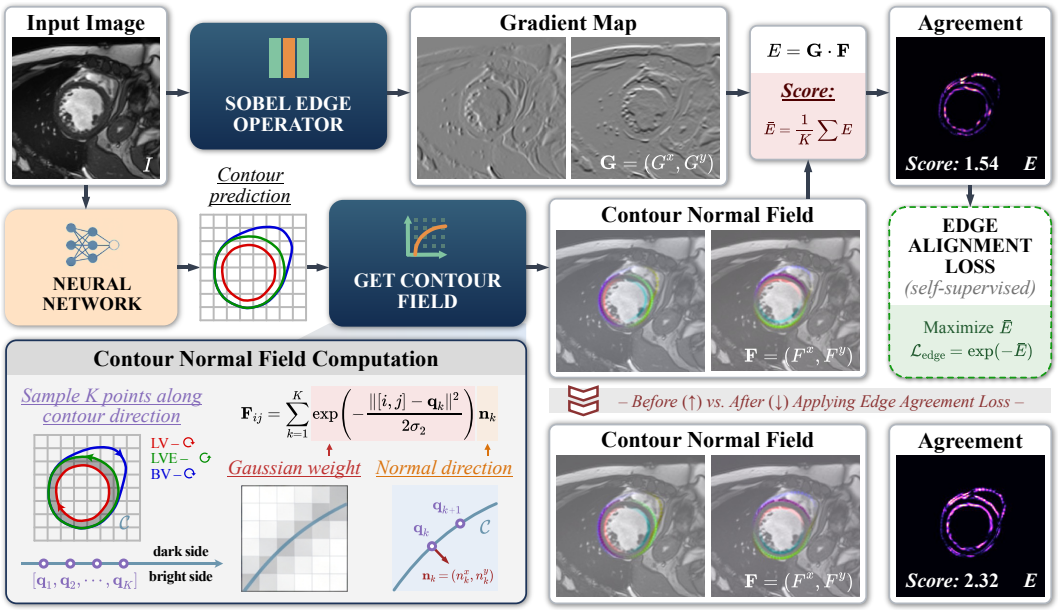

Figure 4: Image–contour edge alignment. Predicted contour normals are aligned with Sobel image gradients, with polarity determined by expected CMR intensity transitions between blood pool, myocardium, and surrounding tissue. This self-supervised loss encourages contours to follow coherent anatomical edges on unlabeled images and enables training with unlabeled data.

where σ controls the Gaussian kernel, set to 1 by default. This formulation propagates the contour normals to nearby pixels, producing a smooth normal field that reflects the local geometry of the predicted contour, with diminishing contributions with increasing distance.

Optimizing Edge Alignment. We define the edge agreement map as the pointwise inner product between the image gradient field and the contour normal field, as $E = \mathbf{G} \cdot \mathbf{F} = G_x F_x + G_y F_y$, which is high when the predicted contour normal aligns with strong image edges of consistent polarity. The edge agreement score is computed as $\bar{E} = K^{-1}\text{sum}(E)$. Finally, the edge alignment loss is defined as $\mathcal{L}_{\text{edge}} = \exp(-\bar{E})$, which promotes alignment of the predicted contour with coherent image edges while remaining differentiable with respect to the contour parameters.

3.4 Self-Supervised Contour Regularization – Consistency Learning

Consistency learning encourages models to produce coherent predictions across transformed views of the same input [42, 43, 44]. As shown in Fig. 3(e), the input image I is scaled to generate an augmented view $I^{\text{aug}} = \text{scale}(I, \lambda)$, where $\lambda \sim \mathcal{U}(0.8, 1.2)$ and the origin $(0, 0)$ is fixed. Feeding both I and I^{aug} into the network yields two sets of Fourier representations:

$$\mathbf{Z} = \{(c_x, c_y, R_0, \mathbf{c})\}, \quad \mathbf{Z}^{\text{aug}} = \{(c_x^{\text{aug}}, c_y^{\text{aug}}, R_0^{\text{aug}}, \mathbf{c}^{\text{aug}})\}, \tag{6}$$

where $\mathbf{c} = \{c_m\}$ denotes the Fourier coefficients.

We enforce consistency learning by aligning the two sets of predicted parameters:

$$\mathcal{L}_{\text{cons}} = \left\|\lambda\,(c_x, c_y, R_0) - (c_x^{\text{aug}}, c_y^{\text{aug}}, R_0^{\text{aug}})\right\|_1 + \|\mathbf{c} - \mathbf{c}^{\text{aug}}\|_1\,, \tag{7}$$

where the location and the average radius (c_x, c_y, R_0) are scaled by λ to match the augmented prediction, while the Fourier coefficients $\mathbf{c} = \{c_m\}$ representing the shape remain invariant to scaling, helping the network capture the relative scale of contours.

3.5 Learning from Multi-Source Supervision

Our framework supports unified learning from mask-labeled, volume-labeled, and unlabeled CMR data. For mask-labeled data, we supervise contour presence classification ($\mathcal{L}_{\text{cls}}$), Fourier coefficient regression ($\mathcal{L}_{\text{coef}}$), and reconstructed contour alignment ($\mathcal{L}_{\text{cont}}$). Specifically, $\mathcal{L}_{\text{cls}}$ predicts whether

Table 1: Comparison with segmentation- and regression-based methods on three public datasets and one in-house cohort. Our contour predictions are rasterized into masks to compute Dice and NSD.

Methods	Dice (%)			NSD (%)			Volume Difference (mL)		
	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
▼ *ACDC Dataset*									
EMCAD† [47]	**96.02**	**89.68**	90.65	92.43	93.81	86.74	-	-	-
nnFormer [17]	95.65	89.58	90.94	92.91	94.18	87.49	-	-	-
MECardNet† [24]	93.58	88.48	**91.60**	91.49	93.95	**88.92**	**4.99** $\pm$ 4.05	8.31 $\pm$ 7.70	7.88 $\pm$ 8.13
CineMA-SEG[1] [4]	93.56	87.93	91.06	-	-	-	5.07	7.53	9.30
nnU-Net [15]	93.43	88.38	91.12	90.97	93.82	87.76	5.76 $\pm$ 4.46	8.36 $\pm$ 6.31	8.20 $\pm$ 8.70
Germain et al. [7]	-	-	-	-	-	-	13.25 $\pm$ 15.77	13.87 $\pm$ 13.55	21.32 $\pm$ 20.37
CineMA-REG [4]	-	-	-	-	-	-	18.42 $\pm$ 20.03	24.26 $\pm$ 24.32	16.71 $\pm$ 19.13
Ours (ResNet)	93.01	88.36	87.70	93.89	95.51	86.37	5.56 $\pm$ 4.33	6.90 $\pm$ 6.52	**7.77** $\pm$ 7.24
Ours (ConvNeXt)	94.17	89.05	87.82	**95.81**	**96.99**	87.56	5.24 $\pm$ 4.50	**6.31** $\pm$ 5.54	7.86 $\pm$ 9.13
▼ *M&Ms Dataset*									
Ma [48]	91.30	**86.27**	**89.37**	86.47	92.08	87.13	-	-	-
Liu et al. [49]	**92.25**	83.78	88.36	87.21	90.74	86.28	-	-	-
MECardNet† [24]	91.43	85.23	88.50	87.04	91.98	86.25	8.16 $\pm$ 6.91	12.00 $\pm$ 10.00	10.15 $\pm$ 10.16
CineMA-SEG [4]	90.70	83.35	87.80	-	-	-	8.54	11.57	9.29
nnU-Net [15]	90.93	84.76	88.75	85.81	91.29	**86.66**	8.23 $\pm$ 6.39	13.65 $\pm$ 9.68	9.49 $\pm$ 10.78
Germain et al. [7]	-	-	-	-	-	-	13.12 $\pm$ 21.22	19.02 $\pm$ 21.84	15.81 $\pm$ 13.63
CineMA-REG [4]	-	-	-	-	-	-	14.49 $\pm$ 15.85	14.98 $\pm$ 14.25	19.46 $\pm$ 24.04
Ours (ResNet)	90.70	84.05	85.37	88.11	92.52	84.98	7.97 $\pm$ 8.70	10.28 $\pm$ 8.45	9.15 $\pm$ 10.07
Ours (ConvNeXt)	91.35	84.78	86.09	**89.06**	**93.19**	86.00	**7.86** $\pm$ 8.63	**9.99** $\pm$ 8.49	**9.04** $\pm$ 9.90
▼ *M&Ms-2 Dataset*									
CAMS-Net† [50]	**94.45**	**91.87**	86.21	90.82	93.61	84.18	-	-	-
MECardNet† [24]	93.86	85.96	**91.14**	89.30	93.23	85.88	7.31 $\pm$ 6.83	7.95 $\pm$ 6.02	13.57 $\pm$ 11.45
CineMA-SEG [4]	93.12	84.80	89.72	-	-	-	7.39	**7.33**	12.62
nnU-Net [15]	93.65	85.34	**91.14**	89.23	92.73	85.80	7.45 $\pm$ 6.11	8.09 $\pm$ 5.95	12.53 $\pm$ 11.26
Germain et al. [7]	-	-	-	-	-	-	14.00 $\pm$ 12.86	14.78 $\pm$ 11.17	19.07 $\pm$ 17.69
CineMA-REG [4]	-	-	-	-	-	-	16.08 $\pm$ 16.17	22.45 $\pm$ 22.09	18.12 $\pm$ 13.42
Ours (ResNet)	93.49	85.28	88.09	91.34	94.12	84.08	7.28 $\pm$ 6.86	7.94 $\pm$ 6.51	11.97 $\pm$ 11.07
Ours (ConvNeXt)	93.85	85.76	88.73	**92.34**	**94.40**	**85.99**	**6.87** $\pm$ 6.55	7.74 $\pm$ 6.49	**11.06** $\pm$ 11.71
▼ *ACP Dataset*									
nnU-Net [15]	94.03	86.31	**89.86**	91.06	92.09	85.67	8.02 $\pm$ 7.57	10.67 $\pm$ 7.65	9.67 $\pm$ 8.98
Ours (ResNet)	95.13	88.53	87.66	95.21	96.29	86.11	5.14 $\pm$ 5.19	**6.29** $\pm$ 5.55	**7.76** $\pm$ 7.69
Ours (ConvNeXt)	**95.65**	**89.23**	87.99	**96.06**	**96.59**	**87.11**	**4.71** $\pm$ 5.05	7.03 $\pm$ 6.60	7.85 $\pm$ 7.83

[1] CineMA is pretrained on 15 million CMR images from 74,916 subjects, about 40$\times$ more data than ours.
† 2D-based methods, running segmentation slice by slice.
* Green shading: segmentation-based methods. Red shading: regression-based methods.

each anatomical contour exists on a slice, $\mathcal{L}_{\text{coef}}$ supervises the Fourier contour parameters, and $\mathcal{L}_{\text{cont}}$ directly constrains the reconstructed contour in image space. For volume-labeled data, we introduce a measurement loss ($\mathcal{L}_{\text{meas}}$), which supervises analytically computed cardiac volumes derived from the predicted contour representations without requiring dense annotations.

For unlabeled data, we use image–contour edge alignment ($\mathcal{L}_{\text{edge}}$) and scale-equivariant contour consistency learning ($\mathcal{L}_{\text{cons}}$) to regularize the contour directly via the clues in the image. Mathematical details of the losses and the final optimization objective are provided in App. B.

4 Experiments

Experiments are conducted on NVIDIA A100 GPUs using PyTorch 2.6.0 [45]. The model is trained with AdamW [46] for 360 epochs, using an initial learning rate of 1e-4, weight decay of 1e-4, batch size 8 (4 for labeled and 4 for unlabeled), 500-step warm-up, followed by a $\times 0.1$ learning-rate decay at epochs 240 and 330. Data augmentation includes in-plane affine transformations with random rotation, translation, scaling, and flipping. Training takes approximately 4 hours on 8 GPUs. Model performance is evaluated using Dice score for overlap accuracy, normalized surface Dice (NSD) for boundary agreement, and absolute volume difference for clinical volume estimation.

Table 2: VLM integration for question answering, with volume difference (mL) reported.

Methods	LV	MYO	RV
MedGemma 1.5 [52]	67.15	38.47	96.46
Qwen3-VL (FT) [53]	14.99	19.10	19.23
Ours + Qwen3-VL [53]	**14.76**	**16.01**	**17.52**

Table 3: Ablation with the number of coefficients.

M	Dice (%)			NSD (%)		
	LV	MYO	RV	LV	MYO	RV
4	93.01	88.36	87.70	93.89	95.51	86.37
6	93.01	88.20	87.88	94.00	95.64	88.03
8	92.67	86.44	87.29	93.45	95.01	87.39

Table 4: Computational efficiency.

Methods	# Param.	FLOPs	Infer. Time
nnU-Net [15]	44.3M	400.3G	189.5 ms
Ours (ResNet)	**42.1M**	**137.4G**	**23.2 ms**

Table 5: Downstream VLM integrated clinical task.

Encoders	LVEF Err.	RVEF Err.	Accuracy
Qwen3-VL	9.10	12.27	0.675
Ours (ResNet)	**6.70**	**9.08**	**0.850**

Datasets. We train a unified model by combining multiple datasets and evaluate it separately on the held-out test splits of each dataset. **1) ACDC [1].** It contains 150 short-axis cine CMR scans, 100 for training and 50 for testing. ED and ES frames are annotated with masks. **2) M&Ms [2].** A multi-center, multi-vendor cine CMR dataset with ED and ES annotations, comprising 150 labeled training scans, 25 unlabeled training scans, 34 validation scans, and 136 test scans. **3) M&Ms-2 [51].** This multi-center dataset includes 360 short-axis cine CMR scans with ED and ES annotations, with 160 scans held out for testing. **4) RO1K.** An in-house dataset of 1,072 cases with CMR report that provides LV and RV volumes at ED and ES frame, all used for measurement-supervised training. **5) ACP.** It contains 95 in-house cine CMR scans annotated across all 30 frames, collected to evaluate generalization to arbitrary cardiac phases (ACP) beyond ED and ES. All cases are used for testing.

4.1 Main Results

The results on CMR datasets are listed in Table 1. On ACDC, although our method is not designed for dense segmentation, its rasterized contours achieve Dice comparable to the representative segmentation baseline nnU-Net [15] for LV and MYO (93.01/88.36 vs. 93.43/88.38), while RV performance is lower. Despite this, our approach shows clear advantages: **1)** Due to the explicit anatomical continuity and smoothness enforced by the contour representation, our predictions exhibit fewer isolated false positives, leading to consistently higher NSD (LV/MYO: 93.89/95.51 vs. 90.97/93.82). **2)** The differentiable volume regression yields more accurate volume estimates (MYO/RV: 6.90/7.77 vs. 8.36/8.20). Replacing the backbone with ConvNeXt [40] further improves performance, increasing Dice by 0.66% and NSD by 1.53% on average, while slightly reducing volume error overall.

On M&Ms and M&Ms-2, we observe a similar trend. While 2D segmentation methods generally obtain higher Dice coefficient, they often exhibit lower NSD, likely because slice-by-slice prediction makes it difficult to maintain smooth and anatomically consistency across slices. In contrast, our method consistently achieves higher NSD for LV and MYO and more accurate volume estimation, reflecting the benefit of the contour learning. ACP serves as an external validation benchmark, where our method outperforms nnU-Net on all metrics except RV Dice, suggesting that our model generalizes beyond ED/ES frames and remains effective across the entire cardiac cycle.

4.2 Integration with VLM

We use our trained model as the vision encoder, feeding its attended anatomical tokens $\mathbf{X}' \in \mathbb{R}^{S\times Q\times C}$ into Qwen3-VL for fine-tuning (see App. C). As shown in Table 2, results on the ACDC dataset demonstrate that the model can estimate volumes through question answering. Although less accurate than explicit Fourier decoding, it outperforms both directly fine-tuned Qwen3-VL and zero-shot MedGemma [52] in volume estimation.

We further explore the potential of our representation for downstream VLM-based clinical reasoning by providing Qwen3-VL with anatomical tokens and Fourier-derived volumes from both ED and ES frames. This evaluation on ACDC considers two question-answering tasks: estimating left and right ventricular ejection fractions (LVEF and RVEF) and identifying cardiac conditions among four categories: normal (NOR), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), and abnormal right ventricle (ARV). Table 5 reports the mean absolute errors of LVEF and RVEF

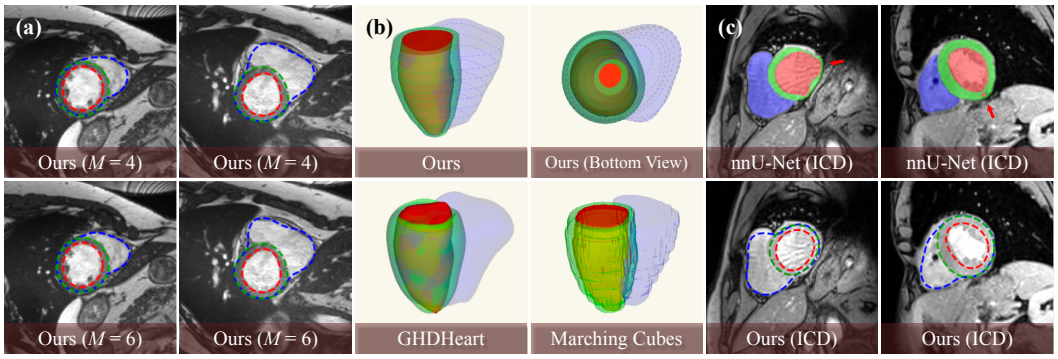

Figure 5: Additional analyses. **(a)** Ablation with the number of coefficients. **(b)** Our method inherently supports cardiac surface reconstruction. GHDHeart [54] and Marching Cubes results are reconstructed from ground-truth mask. **(c)** Robustness to ICD cases with severe artifacts.

Table 6: Ablation with the incremental addition of modules. "*CSA*": Cross-slice attention module.

Modules					**Dice (%)**			**Volume Difference (mL)**		
CSA	$\mathcal{L}_{\text{cont}}$	$\mathcal{L}_{\text{meas}}$	$\mathcal{L}_{\text{cons}}$	$\mathcal{L}_{\text{edge}}$	LV	MYO	RV	LV	MYO	RV
					90.14	82.81	81.01	8.20 ± 6.11	12.13 ± 9.49	11.50 ± 12.99
✓					90.61	83.40	84.62	6.93 ± 5.66	9.24 ± 8.91	11.33 ± 10.89
✓	✓				91.80	86.55	85.28	6.54 ± 4.42	9.00 ± 7.21	10.73 ± 8.50
✓	✓	✓			91.42	86.13	85.04	5.94 ± 4.47	7.64 ± 6.88	8.97 ± 8.67
✓	✓	✓	✓		92.82	87.38	87.55	5.89 ± 4.30	6.94 ± 6.82	8.64 ± 7.93
✓	✓	✓	✓	✓	93.01	88.36	87.70	5.56 ± 4.33	6.90 ± 6.52	7.77 ± 7.24

estimates and disease classification accuracy. These preliminary results suggest the potential of our representation to support VLM-based cardiac functional assessment and diagnostic reasoning.

4.3 More Results

Computational Efficiency. Table 4 evaluates nnU-Net and our method on the same $10 \times 256 \times 256$ input under the same inference environment. Our model uses approximately 1/3 of the FLOPs and is 8.2x faster in this evaluation.

Cardiac Surface Reconstruction. As shown by the M&Ms example in Fig. 5(b), our method predicts analytic contours with ordered point correspondences, enabling direct mesh construction by connecting corresponding points across adjacent contours (see App. D). This is an inherent geometric property of our representation, allowing smooth cardiac surface visualization without the voxel discretization artifacts associated with mask-based methods such as marching cubes). Compared with existing methods such as GHDHeart [54], which deforms predefined mesh templates based on ground truth, our method is template-free, enabling more flexible and faithful reconstruction.

Robustness to Noise and Occlusion. Implantable cardioverter defibrillator (ICD) is a small, battery-powered device placed below the collarbone to monitor heart rhythms, which causes significant noise and occlusion in CMR. We collect a few scans to evaluate robustness under such conditions. As shown in Fig. 5(c), our contour predictions better preserve anatomical structure in the presence of noise and occlusion, demonstrating improved robustness.

4.4 Ablation Studies

Number of Fourier Coefficients. Ablations are performed on ACDC. Table 3 shows the effect of the truncation order M. Increasing M from 4 to 6 produces similar Dice scores across all structures, while slightly improving boundary agreement, especially for RV (visible in Fig. 5(a)), where NSD increases from 86.37% to 88.03%. Since increasing M to 6 provides little gain and further increasing it to 8 plays a negative role, we use $M = 4$ by default as an accuracy–efficiency trade-off.

Effectiveness of Model Components. Table 6 ablates the incremental components of the proposed framework. Adding CSA improves Dice by 1.56%, highlighting the benefit of inter-slice interaction.

Incorporating contour loss $\mathcal{L}_{\text{cont}}$ further improves both segmentation and volume estimation. Introducing volume regression $\mathcal{L}_{\text{meas}}$ substantially reduces volume error (e.g., MYO: 9.00 mL to 7.64 mL; RV: 10.73 mL to 8.97 mL). The scale consistency and edge alignment losses provide additional gains, with the full model achieving the best balance between boundary accuracy and volume estimation.

5 Conclusion

We propose Fourier Contour Learning, a compact and traceable representation for CMR quantification that bridges dense segmentation and direct regression. Building upon this representation, we develop a simple yet effective DETR-style set prediction framework for anatomical contour prediction, enabling unified learning from multi-source supervision, including mask-labeled, measurement-labeled, and unlabeled data. Extensive experiments support three main findings: **1)** The proposed method achieves performance on par with representative segmentation approaches for LV and MYO, while remaining competitive for RV. **2)** By enforcing anatomical continuity and smoothness, the contour representation improves boundary agreement, as reflected by consistently higher NSD for major structures. **3)** The differentiable volume formulation reduces clinically relevant volume error, while the learned contour tokens further provide a compact interface for VLM-based reasoning and surface reconstruction.

Limitations. 1) Some recent dense segmentation models achieve higher Dice scores, especially when evaluated purely as segmentation systems. **2)** The contour-based formulation is particularly well-suited for regular, round structures such as LV and MYO, but remains less expressive for the complex RV geometry. **3)** While we strive for fair comparisons, achieving complete fairness in training data is inherently difficult: Our method can leverage volume-labeled and unlabeled data, whereas competing segmentation methods cannot directly benefit from such supervision.

Future Direction. Volume estimation is not the only goal of designing this framework. Current VLMs still struggle with 4D cine CMR, mainly due to the large number of image tokens required to represent the full spatiotemporal sequence. Our anatomical tokens provide a potential solution by compressing the cine into a compact set of structure-aware representations, while preserving traceable contour-based visualization. Beyond volume estimation, such tokens may serve as an efficient interface for future language-based cardiac reasoning and diagnosis, enabling models to analyze cardiac anatomy, motion, and function with reduced token burden and improved interpretability.

Broader Impact. This work may contribute to more accessible and reliable CMR analysis, potentially improving clinical workflow efficiency, cardiovascular research, and the development of future medical AI systems. Potential risks include inaccurate predictions under unseen imaging conditions, requiring prospective validation and appropriate human oversight before clinical deployment.

Acknowledgments and Disclosure of Funding

Funding. The authors received no external funding or support for this work.

Competing Interests. The authors declare no competing interests.

References

[1] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[2] Victor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021.

[3] Yi Yu, Yixuan Liu, Ziyu Zhang, Parker Martin, Zhenyu Bu, Yuchi Han, and Yuan Xue. Automatic lv localization and short-axis plane estimation from arbitrary cmr slice. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2026.

[4] Yunguan Fu, Wenjia Bai, Weixi Yi, Charlotte Manisty, Anish N Bhuva, Thomas A Treibel, James C Moon, Matthew J Clarkson, Rhodri Huw Davies, and Yipeng Hu. A versatile foundation model for cine cardiac magnetic resonance image analysis tasks. *arXiv preprint arXiv:2506.00679*, 2025.

[5] Avan Suinesiaputra, Brett R Cowan, Ahmed O Al-Agamy, Mustafa A Elattar, Nicholas Ayache, Ahmed S Fahmy, Ayman M Khalifa, Pau Medrano-Gracia, Marie-Pierre Jolly, Alan H Kadish, et al. A collaborative resource to build consensus for automated left ventricular segmentation of cardiac mr images. *Medical Image Analysis*, 18(1):50–62, 2014.

[6] Sulaiman Vesal, Mingxuan Gu, Andreas Maier, and Nishant Ravikumar. Spatio-temporal multi-task learning for cardiac mri left ventricle quantification. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2698–2709, 2020.

[7] Philippe Germain, Aissam Labani, Armine Vardazaryan, Nicolas Padoy, Catherine Roy, and Soraya El Ghannudi. Segmentation-free estimation of left ventricular ejection fraction using 3d cnn is reliable and improves as multiple cardiac mri cine orientations are combined. *Biomedicines*, 12(10):2324, 2024.

[8] Steffen E. Petersen, Paul M. Matthews, Jane M. Francis, Matthew D. Robson, Filip Zemrak, Redha Boubertakh, Alistair A. Young, Sarah Hudson, Peter Weale, Steve Garratt, Rory Collins, Stefan Piechnik, and Stefan Neubauer. Uk biobank's cardiovascular magnetic resonance protocol. *Journal of Cardiovascular Magnetic Resonance*, 18(1):8, 2016.

[9] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. In *Conference on Neural Information Processing Systems*, 2025.

[10] Yifan Shen, Jian Xu, Boyi Li, Yuner Zhang, Tianjiao Yu, Bingxuan Li, Houze Yang, Rushi Wang, and Xu Cao. Chronovision: Temporal reasoning via latent state reconstruction. *arXiv preprint arXiv:2608.05631*, 2026.

[11] Taiping Qu, Hongkai Zhang, Lantian Zhang, Can Zhao, Nan Zhang, Hui Wang, Zhen Zhou, Mingye Zou, Kairui Bo, Pengfei Zhao, Xingxing Jin, Zixian Su, Kun Jiang, Huan Liu, Yu Du, Maozhou Wang, Ruifang Yan, Zhongyuan Wang, Tiejun Huang, Lei Xu, and Henggui Zhang. Baai cardiac agent: An intelligent multimodal agent for automated reasoning and diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging. *arXiv Preprint arXiv:2604.04078*, 2026.

[12] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026.

[13] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Yuanzhe Liu, and Ismini Lourentzou. Core3d: Collaborative reasoning as a foundation for 3d intelligence. *arXiv preprint arXiv:2512.12768*, 2025.

[14] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Onkar Susladkar, Yuanzhe Liu, Xiaona Zhou, and Ismini Lourentzou. Elsa3d: Elastic semantic anchoring for unified 3d understanding and generation. *arXiv preprint arXiv:2607.06565*, 2026.

[15] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[16] 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*, pages 234–241, 2015.

[17] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 32:4036–4045, 2023.

[18] Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[19] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*, pages 205–218, 2022.

[20] Ziyu Zhang, Yi Yu, and Yuan Xue. Rethinking roi strategy in interactive 3d segmentation for medical images. In *CVPR 2025 Challenge on Foundation Models for 3D Biomedical Image Segmentation*, 2025.

[21] Ziyu Zhang, Yi Yu, and Yuan Xue. Unlocking 2d promptable foundation models for 3d vessel segmentation by automatic prompt generation. In *International Conference on Medical Imaging with Deep Learning*, 2026.

[22] Ziyu Zhang, Yi Yu, Simeng Zhu, Ahmed Aly, Yunhe Gao, Ning Gu, and Yuan Xue. Atlas is your perfect context: One-shot customization for generalizable foundational medical image segmentation. In *European Conference on Computer Vision*, 2026.

[23] Guoqiang Ren, Zhen Chen, Pengxiang Su, Da Li, Xiaoping Yang, Di Gai, Xin Wei, Weifeng Xu, Hongping Chen, Xiaoguang Zhao, et al. Cfg-mambanet: Contextual and frequency-guided mamba network for medical image segmentation. *npj Digital Medicine*, 2026.

[24] Hamed Aghapanah, Reza Rasti, Faezeh Tabesh, Hamidreza Pouraliakbar, Hamid Sanei, and Saeed Kermani. Mecardnet: A novel multi-scale convolutional ensemble model with adaptive deep supervision for precise cardiac mri segmentation. *Biomedical Signal Processing and Control*, 100:106919, 2025.

[25] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[26] Fangzhou Liao, Xi Chen, Xiaolin Hu, and Sen Song. Estimation of the volume of the left ventricle from mri images using deep neural networks. *IEEE Transactions on Cybernetics*, 49(2):495–504, 2017.

[27] Tianyang Li, Benzheng Wei, Jinyu Cong, Yanfei Hong, and Shuo Li. Direct estimation of left ventricular ejection fraction via a cardiac cycle feature learning architecture. *Computers in Biology and Medicine*, 118:103659, 2020.

[28] Gongning Luo, Wei Wang, Clara Tam, Kuanquan Wang, Shaodong Cao, Henggui Zhang, Bo Chen, and Shuo Li. Dynamically constructed network with error correction for accurate ventricle volume estimation. *Medical Image Analysis*, 64:101723, 2020.

[29] Manuel Pérez-Pelegrí, José V Monmeneu, María P López-Lereu, Lucía Pérez-Pelegrí, Alicia M Maceira, Vicente Bodí, and David Moratal. Automatic left ventricle volume calculation with explainability through a deep learning weak-supervision methodology. *Computer Methods and Programs in Biomedicine*, 208:106275, 2021.

[30] Athira J Jacob, Indraneel Borgohain, Teodora Chitiboi, Puneet Sharma, Dorin Comaniciu, and Daniel Rueckert. Towards a cardiovascular magnetic resonance foundation model for multi-task cardiac image analysis. *Journal of Cardiovascular Magnetic Resonance*, 27(2):101967, 2025.

[31] Yundi Zhang, Chen Chen, Suprosanna Shit, Sophie Starck, Daniel Rueckert, and Jiazhen Pan. Whole heart 3d+ t representation learning through sparse 2d cardiac mr images. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 359–369, 2024.

[32] Yundi Zhang, Paul Hager, Che Liu, Suprosanna Shit, Chen Chen, Daniel Rueckert, and Jiazhen Pan. Towards cardiac mri foundation models: Comprehensive visual-tabular representations for whole-heart assessment and beyond. *Medical Image Analysis*, page 103756, 2025.

[33] Tong Zhang, Yin Zhuang, Guanqun Wang, He Chen, Lianlin Li, and Jun Li. A unified remote sensing object detector based on fourier contour parametric learning. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–25, 2025.

[34] Eric Upschulte, Stefan Harmeling, Katrin Amunts, and Timo Dickscheid. Contour proposal networks for biomedical instance segmentation. *Medical Image Analysis*, 77:102371, 2022.

[35] Jiawei Cao, Chaochen Gu, Hao Cheng, Xiaofeng Zhang, Kaijie Wu, and Changsheng Lu. Efdtr: Learnable elliptical fourier descriptor transformer for instance segmentation. In *International Conference on Machine Learning*, 2025.

[36] Mehmet Bahadir Erden, Selahattin Cansiz, Onur Caki, Haya Khattak, Durmus Etiz, Melek Cosar Yakar, Kerem Duruer, Berke Barut, and Cigdem Gunduz-Demir. Fourierloss: Shape-aware loss function with fourier descriptors. *Neurocomputing*, 638:130155, 2025.

[37] Jia Shen, Jun Zhang, Xueyu Liu, Yunfei Zhang, Guangze Shi, Feixue Shao, Hangbei Cheng, and Yongfei Wu. Fourier transform-based shape constrained framework for generalizable medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 295–309. Springer, 2025.

[38] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[40] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[41] Yi Yu, Botao Ren, Peiyuan Zhang, Mingxin Liu, Junwei Luo, Shaofeng Zhang, Feipeng Da, Junchi Yan, and Xue Yang. Point2rbox-v2: Rethinking point-supervised oriented object detection with spatial layout among instances. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19283–19293, 2025.

[42] Yi Yu, Xue Yang, Qingyun Li, Yue Zhou, Feipeng Da, and Junchi Yan. H2rbox-v2: Incorporating symmetry for boosting horizontal box supervised oriented object detection. In *Conference on Neural Information Processing Systems*, 2023.

[43] Yi Yu, Xue Yang, Qingyun Li, Feipeng Da, Jifeng Dai, Yu Qiao, and Junchi Yan. Point2rbox: Combine knowledge from synthetic visual patterns for end-to-end oriented object detection with single point supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16783–16793, 2024.

[44] Yi Yu, Xue Yang, Yansheng Li, Zhenjun Han, Feipeng Da, and Junchi Yan. Wholly-wood: Wholly leveraging diversified-quality labels for weakly-supervised oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4438–4454, 2025.

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems*, volume 32, pages 8024–8035, 2019.

[46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[47] Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11769–11779, 2024.

[48] Jun Ma. Histogram matching augmentation for domain adaptation with application to multi-centre, multi-vendor and multi-disease cardiac image segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 177–186. Springer, 2020.

[49] Yiwei Liu, Liang Zhong, Lingyi Wen, and Yuankai Wu. Uncertainty-based ensemble learning in cmr semantic segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8057–8061, 2026.

[50] Abbas Khan, Muhammad Asad, Martin Benning, Caroline Roney, and Gregory Slabaugh. Cams: convolution and attention-free mamba-based cardiac image segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1893–1903, 2025.

[51] Carlos Martín-Isla, Víctor M. Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J. Fulton, et al. Deep learning segmentation of the right ventricle in cardiac mri: The m&ms challenge. *IEEE Journal of Biomedical and Health Informatics*, 27(7):3302–3313, 2023.

[52] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[53] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv Preprint arXiv:2511.21631*, 2025.

[54] Yihao Luo, Dario Sesia, Fanwen Wang, Yinzhe Wu, Wenhao Ding, Kamrul Hasan, Jiahao Huang, Fadong Shi, Anoop Shah, Amit Kaura, Jamil Mayet, Guang Yang, and Choon Hwai Yap. Explicit differentiable slicing and global deformation for cardiac mesh reconstruction. *Medical Image Analysis*, 111:103999, 2026.

A Module Design Details

Cross-Slice Attention. Given a stack of S slices, the decoder produces query features $\mathbf{X} \in \mathbb{R}^{S\times Q\times C}$, where $Q = 3$ denotes the number of curve queries and $C = 256$ is the token dimension. To capture inter-slice contextual information, we apply a lightweight cross-slice interaction module that performs self-attention along the slice dimension independently for each query stream.

Specifically, the input tokens are first normalized with layer normalization, $\widetilde{\mathbf{X}} = \mathrm{LN}(\mathbf{X})$. Multi-head self-attention with 8 heads is then applied over the slice axis, treating anatomical queries as parallel streams. The attention output can be written as:

$$\mathrm{Attn}(\widetilde{\mathbf{X}}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{B}\right)\mathbf{V}, \tag{8}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are the query, key, and value projections of $\widetilde{\mathbf{X}}$, and $d = 32$ denotes the per-head feature dimension. $\mathbf{B} \in \mathbb{R}^{S\times S}$ is a distance-aware bias with $B_{ij} = -\beta|i - j|$ and $\beta = 0.25$, which encourages stronger interaction between nearby slices.

The attended features are then further refined by a feed-forward network with expansion ratio 2:

$$\mathrm{FFN}(\cdot) = W_2\,\mathrm{GELU}(W_1\,\mathrm{LN}(\cdot)). \tag{9}$$

The feed-forward output is added to the attention output, and then the input features are updated through a residual connection with learnable scale α:

$$\mathbf{X}' = \mathbf{X} + \alpha\left(\mathrm{Attn}(\widetilde{\mathbf{X}}) + \mathrm{FFN}\Big(\mathrm{Attn}(\widetilde{\mathbf{X}})\Big)\right), \tag{10}$$

where α is initialized to 0, so the module starts as an identity mapping and gradually learns how much cross-slice interaction to inject. The resulting $\mathbf{X}' \in \mathbb{R}^{S\times Q\times C}$ is then passed to the downstream Fourier contour head for contour prediction or the vision-language model for question answering.

Fourier Contour Head. The Fourier contour head consists of two MLP branches: a coefficient branch and a classification branch. The coefficient branch is a two-hidden-layer MLP with ReLU activations, producing $2M + 3$ outputs, where the first two terms, corresponding to the normalized contour center (c_x, c_y), are constrained to $(0, 1)$ by a sigmoid. The third term, corresponding to the base radial scale R_0, is enforced to be positive through an exponential transform, scaled by a constant factor (0.05 in our implementation). The remaining $2M$ higher-order Fourier coefficients are bounded to $(-0.2, 0.2)$ using a shifted sigmoid. In parallel, a separate one-hidden-layer MLP predicts a scalar classification logit p, indicating whether the queried structure is present. This parameterization stabilizes optimization by constraining the low-frequency geometric terms to anatomically plausible ranges while allowing the higher-order coefficients to model local contour deformation.

B Loss Function Details

We train the model with a combination of contour classification, coefficient regression, contour reconstruction, and volumetric consistency losses. Let $p^{q,s}$ denote the predicted contour-presence logit for structure query q on slice s, with binary target $y^{q,s} \in \{0, 1\}$. Let $\mathbf{z}^{q,s} \in \mathbb{R}^{2M+3}$ and $\mathbf{z}^{q,s}_{\mathrm{GT}} \in \mathbb{R}^{2M+3}$ denote the predicted and ground-truth Fourier contour parameters:

$$\mathbf{z} = (c_x,\, c_y,\, R_0,\, \mathrm{Re}(c_1),\, \mathrm{Im}(c_1),\, \ldots,\, \mathrm{Re}(c_M),\, \mathrm{Im}(c_M))\,. \tag{11}$$

The contour-presence loss is a binary cross-entropy loss:

$$\mathcal{L}_{\mathrm{cls}} = \frac{1}{|\Omega|}\sum_{(q,s)\in\Omega} \mathrm{BCEWithLogits}(p^{q,s}, y^{q,s}), \tag{12}$$

where Ω denotes the set of all query-slice pairs. This term supervises whether each anatomical contour is present on a given slice.

For positive samples only, we further supervise the Fourier parameters. Denoting by $\Omega^+ = \{(q, s) \in \Omega \mid y^{q,s} = 1\}$ the set of positive query-slice pairs, the coefficient regression loss is defined as:

$$\mathcal{L}_{\mathrm{coef}} = \frac{1}{|\Omega^+|}\sum_{(q,s)\in\Omega^+} \left\|\mathbf{z}^{q,s} - \mathbf{z}^{q,s}_{\mathrm{GT}}\right\|_1, \tag{13}$$

which encourages accurate recovery of the global Fourier shape parameterization, including the contour center, base scale, and higher-order harmonic coefficients.

We also supervise the reconstructed boundary by converting both predicted and ground-truth coefficients into sampled contours with K points, expressed as $\Gamma(\mathbf{z})$. The contour loss is:

$$\mathcal{L}_{\text{cont}} = \frac{1}{|\Omega^+|} \sum_{(q,s)\in\Omega^+} \left\| \Gamma(\mathbf{z}^{q,s}) - \Gamma(\mathbf{z}^{q,s}_{\text{GT}}) \right\|_1 , \tag{14}$$

which directly penalizes boundary misalignment in image space and therefore provides more explicit supervision on the contour shape.

Finally, we impose global measurement through a volume-based loss. Let $A^{q,s}$ denotes the enclosed area for a predicted Fourier representation, calculated analytically as in Eq. (4). The contribution of each slice is weighted by the predicted contour-presence score, and the cumulative volume estimate for structure query q is obtained by summing across slices:

$$V^q = \sum_s \text{sigmoid}(p^{q,s})\, A^{q,s}. \tag{15}$$

Since the predicted contours correspond to the LV endocardial surface (LV), LV epicardial surface (LVE), and biventricular surface (BV), the final regional volumes are derived by subtraction to match the standard dataset definitions:

$$\mathbf{V} = \left[V^{LV},\ V^{MYO},\ V^{RV}\right] = \left[V^{LV},\ V^{LVE} - V^{LV},\ V^{BV} - V^{LVE}\right]. \tag{16}$$

The measurement loss is defined with predicted and ground-truth volume as $\mathcal{L}_{\text{meas}} = \|\mathbf{V} - \mathbf{V}_{\text{GT}}\|_1$, encouraging contour predictions to remain globally consistent with measurement supervision.

Finally, we train the model using a weighted combination of complementary objectives:

$$\mathcal{L} = w_{\text{cls}}\mathcal{L}_{\text{cls}} + w_{\text{edge}}\mathcal{L}_{\text{edge}} + w_{\text{cont}}\mathcal{L}_{\text{cont}} + w_{\text{coef}}\mathcal{L}_{\text{coef}} + w_{\text{meas}}\mathcal{L}_{\text{meas}} + w_{\text{cons}}\mathcal{L}_{\text{cons}}, \tag{17}$$

where $\mathcal{L}_{\text{edge}}$ and $\mathcal{L}_{\text{cons}}$ are described in Sec. 3.3 and Sec. 3.4. $w_{(\cdot)}$ are adjusted to make each supervised loss term around 0.1, and self-supervised loss around 0.01 at the end of training.

C Integration with Qwen3-VL

We integrate our model with a vision-language model (VLM) by using the predicted anatomical tokens as visual inputs. Specifically, the queried features $\mathbf{X}' \in \mathbb{R}^{S\times Q\times C}$, produced by our cross-slice attention, are first linearly projected to match the embedding dimension required by Qwen3-VL-2B [53]. The resulting features are reshaped into $S \times Q$ visual tokens and concatenated with textual tokens as input to the VLM. The entire pipeline, including our model and the VLM, is jointly fine-tuned end-to-end, allowing the anatomical tokens to adapt to the downstream reasoning task.

Compared with the default vision input pipeline of Qwen3-VL [53], or other medical VLMs such as MedGemma [52], this representation significantly reduces the number of visual tokens, providing a compact yet structured anatomical encoding. This efficiency is particularly beneficial for scaling to full 4D cine CMR, enabling future whole-heart reasoning across space and time.

To support quantitative reasoning, voxel size information derived from the affine matrix of the CMR is explicitly included in the prompt. This allows the model to correctly convert voxel counts into physical volumes. We adopt a fixed question template, for example:

> **Q:** In this CMR image, each voxel corresponds to 27 mm^3. What is the LV volume?
> **A:** The LV volume is 131 mL.

This example highlights how our compact anatomical representation can be directly interfaced with a VLM for downstream question answering and quantitative analysis.

D Surface Reconstruction Details

Thanks to our analytic prediction, surface reconstruction can be performed directly from the contour geometry without relying on voxel masks or template fitting. For each 3D structure, the predicted

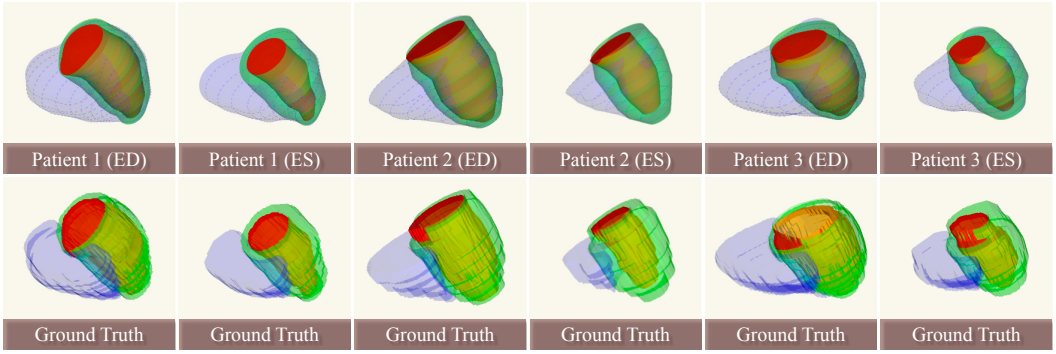

Figure 6: More reconstruction results. The first row shows surfaces reconstructed from our predicted Fourier contours, while the second row shows the corresponding ground truth reconstructed from segmentation masks via marching cubes. Our method preserves consistent structures while producing smoother surfaces. A video of 4D cine reconstruction is provided in the code repository.

Fourier parameters are first sampled using Eq. (1) to obtain a set of contours:

$$\mathcal{C}_s = \{(x_{s,k}, y_{s,k})\}_{k=1}^{K}, \tag{18}$$

where s is the slice index and K is the number of sampled contour points. This gives a consistent ordered point set on every slice, which can then be lifted into 3D vertices using the slice affine.

Afterward, we reconstruct the surface by connecting adjacent contour rings along the slice direction. For two neighboring slices, each quadrilateral strip is split into two triangles to form the surface:

$$T_{s,k}^{(1)} = (\mathbf{p}_{s,k},\, \mathbf{p}_{s+1,k},\, \mathbf{p}_{s+1,k+1}), \quad T_{s,k}^{(2)} = (\mathbf{p}_{s,k},\, \mathbf{p}_{s+1,k+1},\, \mathbf{p}_{s,k+1}), \tag{19}$$

where $\mathbf{p}_{s,k}$ is the k-th 3D contour point on slice s, and the point index wraps around cyclically.

Since each predicted contour represents the center of a slice rather than its boundary, we further interpolate intermediate rings between adjacent slices and extend the first and last contours by half a slice. This allows the reconstructed mesh to span the full volumetric extent. The open ends are then closed by triangulating the first and last rings to form surface caps.

This contour-driven reconstruction yields smooth, template-free, high-quality cardiac surfaces, with more results displayed in Fig. 6, while avoiding the staircase artifacts and discretization errors that commonly arise when surfaces are extracted from voxelized segmentation masks.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the proposed Fourier contour representation, the unified learning framework, and the research scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the conclusion section we summarize three limitations of our method.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not present formal theoretical results, theorems, or proofs. The mathematical expressions in the method section define the Fourier contour parameterization, analytic area computation, and training losses used by the model.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The source code and pretrained model weights are publicly available on GitHub: `https://github.com/yuyi1005/cardiocontour`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The source code and pretrained model weights are released at the following GitHub repository: `https://github.com/yuyi1005/cardiocontour`. The three datasets used to support the main experimental results of this paper are already publicly available online.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The details are specified in the experiment section.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: For our reproduced experiments and proposed models, volume differences are reported as mean ± standard deviation across test cases. For literature baselines where per-case predictions or uncertainty estimates are not available, we report the values provided in the original papers and leave unavailable error bars blank.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The GPU and the training time is stated in the experiment section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presented in this paper focuses on methodological advances. To the best of our knowledge, the work conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: It is discussed in conclusion section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: To the best of our knowledge, the paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The three public datasets used in our research have been cited properly. The ACDC dataset is free of charge under [CC BY-NC-SA 4.0 license]. The M&Ms and M&Ms-2 datasets do not explicitly specified the license, but it mentions "Permission is granted to view and use the Dataset without charge for research purposes only."

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The source code and pretrained model weights have been publicly released through our GitHub repository (`https://github.com/yuyi1005/cardiocontour`). The repository also provides documentation and usage instructions to facilitate reproducibility and further research.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study does not involve crowdsourcing. The in-house data are retrospective CMR studies analyzed under institutional approval, which is addressed in the IRB question.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

Justification: The retrospective use of the in-house CMR datasets was approved by the institutional IRB under a retrospective study protocol. The data were handled in de-identified form for model development and evaluation, with no prospective intervention or participant contact.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.