

Supplemental Material

1 Image Preprocessing and Annotation

The raw data consisted of desensitized MP4 format fetal four-chamber view video files exported from ultrasound machines. Approximately 30 to 60 frames were extracted from each original fetal heart video and stored in JPEG format. Six ultrasound physicians with 5 to 8 years of experience were involved in the annotation of fetal four-chamber view images. Using the open-source tool LabelImg (version 2.0), they annotated the anatomical structures (left ventricle, left atrium, right ventricle, right atrium, apex, mitral valve, foramen ovale, interventricular septum, etc.) with the smallest possible rectangular bounding boxes to train the model to identify the standard four-chamber heart view. Additionally, a self-developed polygonal annotation tool was used to outline the actual boundaries of the anatomical structures to train the model for segmenting these structures, which is essential for calculating cardiac function parameters. During polygonal annotation, the tool provided playback and rewind functions, allowing the annotators to view the preceding and following frames of the annotated frame to assist in accurately identifying the boundaries of anatomical structures that may shift in position within the video, ensuring the accuracy of the annotations. After the initial annotation, two ultrasound physicians with over 10 years of experience reviewed and corrected any inaccurate bounding boxes and polygonal annotations. The reviewed and approved data were then used for training and fine-tuning the deep learning model.

2 Data Augmentation

To enhance the model’s generalization ability for complex scenarios, data augmentation was implemented in offline and online parts:

For offline data augmentation, the operation includes (1) Random Scaling: Using a base scaling ratio of 0.3, the scale distribution of the input images was dynamically adjusted to force the model to learn robust feature representations for multi-scale targets. (2) Random Rotation: Images were rotated within a continuous range of angles to enhance the model’s adaptability to directional changes of the target, reducing detection bias caused by viewpoint differences. (3) Random Contrast Adjustment: By dynamically adjusting the difference between the bright and dark areas of the images, the visual features under varying brightness conditions were simulated to improve the model’s robustness to changes in image brightness.

For online data augmentation, a combination of multi-dimensional geometric transformations and color space perturbations was applied, including (1) Geometric Transformation Enhancement: A composite transformation involving random cropping (30% amplitude in all directions), scaling (0.5 to 1.5 times), translation, and rotation (-15° to $+15^\circ$) was used to simulate the multi-scale spatial distribution characteristics of the target in real-world scenarios. This strategy, through parametric random perturbation, forced the model to learn the feature representations of targets from different viewpoints and scales, alleviating overfitting caused by a single viewpoint. (2) Color Space Perturbation Module: The RandomDistort technique was employed to decouple multi-dimensional color features, including: non-linear perturbation in the HSV space: hue shift ($\pm 30\%$), saturation adjustment ($\pm 50\%$), and brightness fluctuation ($\pm 30\%$); random permutation of RGB channels with a 0.5 probability, the order of the three channels was swapped to break the inherent correlation of color features. This module simulated different imaging environments to enhance the model’s detection robustness under low-contrast and color-biased conditions. (3) Target Integrity Protection Mechanism: A two-stage adaptive sampling strategy was introduced: random expansion with a 0.5 probability, the image was randomly expanded to 1.0 to 4.0 times its size to simulate the presence of the target in different backgrounds; target-aware cropping when cropping resulted in a target box area loss exceeding 50%, a hard negative sample filtering mechanism was triggered to ensure the semantic integrity of the effective target during training, preventing the model from producing erroneous detection results due to overconfidence.

After undergoing the above augmentation techniques, the images were resized to a 640x640 pixel resolution while maintaining the aspect ratio for model training.

3 Model Input and Output

The common human-based workflow for performing fetal cardiac measurement consists of a sequence of steps: (1) identifying a sequence that smoothly displays the standard fetal four-chamber view from the whole screening video; (2) determining the systolic and diastolic phase; (3) manually segmenting and drawing annotations on frames to outline cardiac chambers; and (4) manually referring to standard reference guidelines to calculate the measurement by geometric information of the annotated object, such as area or length. The above data pipeline is executed in sequence, which makes the evaluation of fetal heart function parameters inefficient and manual analysis time-consuming. Furthermore, the limited processing capacity of human operators merely support to focus on a few frames at once. The selection of measurement frames and endpoint positions involves significant subjectivity, leading to poor repeatability of measurements and even failure to replicate results in regions lacking expert resources.

To solve these problem, our AI workflow unifies these steps by a identify first and judge later way. The motivation is to shift the analysis of fetal echocardiography from the selection of key frames to the direct selection of key cardiac structures. The powerful computational capabilities of AI enable parallel analysis of every frame in fetal electrocardiograms, efficiently detecting and delineating potential anatomical boundaries. This significantly enhances data processing efficiency, thereby reducing

evaluation time for fetal cardiac function. Furthermore, the integration of the holistic analysis with the expert knowledge allows the objective automated selection of optimal measurement windows, improving the reproducibility of cardiac function assessments. Most importantly, the powerful learning capacity of AI enables recognizing visual features in complex fetal positions through extensive training with expert-collected and -annotated data. This ensures high-quality structural detection and precise boundary delineation. To this end, the AI workflow automates the entire process without the need for human intervention, greatly improving the efficiency and repeatability of fetal cardiac function evaluation.

4 RT-DETR Detector

In the identification process, the AI workflow rapidly analyses each frame of videos stored on the digital imaging and communications in medicine (DICOM) files of the patient examination and outline the visual structure that is familiar with the cardiac structure of interest. We trained a famous deep learning-based detector named RT-DETR, which combines the advantages of convolutional neural network (CNN) and Transformer. RT-DETR is specifically designed to account for the distinctive characteristics of fetal echocardiography, aiming to enhance accuracy while reducing computational cost. Segmenting fetal cardiac chambers requires capturing global structural information to guide feature extraction within locally similar blood-flow regions. Moreover, due to ultrasound image noise and the relatively large size of the cardiac chambers, there is no need to extract the extremely fine-grained local features required for natural images. RT-DETR employs a Transformer only on the smallest-scale feature map to extract global context. It then uses a lightweight feature pyramid to propagate this global information across all scales, generating uniformly sized feature blocks. Thus, RT-DETR effectively integrates global and local information within a compact architecture while significantly reducing the computational overhead associated with Transformers. The architecture retains convolutional neural network as the backbone network (using ResNet18 as the convolutional neural network backbone) to perform initial feature extraction on the input image, generating several two-dimensional feature maps. This step maps the image to a unified expression space and performs downsampling to reduce computational load before feeding it into the main Transformer module to complete the detection task. To meet the GPU memory requirements for Transformer training, a server with two RTX 4090 graphics cards was used for training, with approximately 200 rounds of convergence and early termination of training after about 3 days. When a fetal four-chamber heart ultrasound image is input into the model, the model outputs the detection boxes and structure labels for all key structures in the section, as shown in Fig. S1.

The loss function contains two parts: (1) Classification Loss:

$$\mathcal{L}_{cls} = -\sum_{i=1}^N [\hat{y}_i \log(p_i) + (1 - \hat{y}_i) \log(1 - p_i)], \quad (1)$$

where $\hat{y}_i$ and p_i denotes the ground true and predicted category, and N is the box number. (2) Bounding Box Regression Loss:

$$\mathcal{L}_{box} = \lambda_1 \sum_{i=1}^4 |b_{pred}^i - b_{gt}^i| + \lambda_2 (1 - GIoU(b_{pred}, b_{gt})), \quad (2)$$

where b_{pred} and b_{gt} denote the center coordinates, width, and height of the predicted and ground-truth box, and $GIoU(\cdot)$ measures the overlap between the predicted box and the ground-truth box, addressing the issue of gradient disappearance in traditional IoU when there is no overlap.

5 Mask-Dino Segmentation Head

We designed a segmentation head to extend RT-DETR, which enables the unified Transformer-based decoding for both the object detection and segmentation tasks. The algorithm inherits most of the improvements from RT-DETR, such as cross attention with anchor box guidance, query selection, and noise reduction training. Therefore, the detection and segmentation tasks can help each other by leveraging shared intermediate features to complete both detection and boundary delineation. The segmentation head reuses the front-end backbone network, adding only about 6-8 ms to the inference time to achieve high-precision real-time detection of cardiac function. During training, the loss is calculated only for the annotated four chambers, cardiac outline, and related structures (since bounding box target detection and instance segmentation have different numbers of annotated objects), thereby enabling joint training of the detection head and segmentation head. The model was trained on a dual RTX 4090 server for 3 days, with approximately 200 rounds of training before termination. The segmentation results of the fetal cardiac ultrasound images by the AI model are shown in Fig. S2.

The loss function contains two parts: (1) Segmentation Loss:

$$\mathcal{L}_{seg} = -\sum_{i=1, j=1}^{H, W} [\hat{y}_{i,j} \log(p_{i,j}) + (1 - \hat{y}_{i,j}) \log(1 - p_{i,j})], \quad (3)$$

where $\hat{y}_i$ and p_i denotes the ground true and predicted label at the position (i, j) , and H, W is the height and weight of the image. (2) Dice Loss:

$$\mathcal{L}_{dice} = 1 - \frac{2 \sum_{i=1, j=1}^{H, W} p_{i,j} \hat{y}_{i,j}}{\sum_{i=1, j=1}^{H, W} p_{i,j} + \sum_{i=1, j=1}^{H, W} \hat{y}_{i,j}}, \quad (4)$$

In summary, the total loss function is formulated as:

$$\mathcal{L}_{loss} = \lambda_{cls} \mathcal{L}_{cls} + \lambda_{box} \mathcal{L}_{box} + \lambda_{seg} \mathcal{L}_{seg} + \lambda_{dice} \mathcal{L}_{dice}, \quad (5)$$

where the factors for each term are set as $\lambda_{cls} = 4$, $\lambda_{box} = 2$, $\lambda_{seg} = 5$, $\lambda_{dice} = 5$. Both the detection and segmentation head are based on the DETR architecture, and the segmentation task can be regarded as a more refined object detection task. Therefore, the same hyperparameters are used for both, with the difference lying in the different

loss functions used and the additional segmentation head in the segmentation model. The training hyperparameters for the detection and segmentation networks are shown in Table S1.

6 Geometric Calculation of Cardiac Function Parameters

Based on the bounding boxes and polygonal contours of the fetal cardiac structures output by the AI model, cardiac function parameters are calculated using geometric, mathematical, and image processing methods. For example, the calculation of cardiac transverse diameter requires determining two geometric endpoints through geometric methods and calculating the distance between them. Each measurement item is obtained through a similar calculation process. The cardiac cycle is calculated based on continuous video frames. The calculation of the cardiac cycle relies on the accurate segmentation of the four chambers by the segmentation head, identifying the key frames of systole and diastole through the dynamic and regular area changes, and finally estimating the fetal cardiac cycle based on these key frames. Since stored images and real-time ultrasound examination videos often contain more than one complete cardiac cycle, the model calculates the measurement values for each complete cardiac cycle in real-time and designs a strategy based on section standard and image quality assessment to determine the optimal cardiac cycle, approximating the optimal judgment standard of ultrasound physicians. The fetal cardiac function parameters measured in the four-chamber heart section using the AI model are listed in Table S2 and S3.

7 Supplemental Experiment Result

Table S1 lists the hardware configuration of the model testing. Table S4 presents the results of five state-of-the-art models and our RT-DERT on the full validation data. Also, the significance test was conducted to verify the Dice metric differences between these five methods and our approach. Fig. S3 shows two cases of the segmentation and measurement results. The segmentation errors primarily arise from two sources: global boundary shifts, and locally ambiguous boundaries, particularly between cardiac chambers and valves during valve opening. Such global shifts do not introduce substantial errors in the final measurements because endpoint errors in measurement points are often consistent in direction and thus tend to cancel each other out in the final calculation. Furthermore, cardiac chamber measurements are prioritized at end-diastole—when valves are clearly closed, therefore visualized—the impact of segmentation uncertainty on the final measurements is further minimized. The human-machine measurement results are listed in Table S5-S8. Fig. S4-S6 show the Bland-Altman graphs for global sphericity index, fractional area change of right ventricle, and ejection fraction of left ventricle in the normal dataset. Fig. S7-S9 show the Bland-Altman graphs for these parameters in the abnormal dataset.

8 Detail of Z-score modeling

Various cardiac parameters exhibit changes throughout the biparietal diameter (BPD), head circumference (HC), abdominal circumference (AC), femur length (FL), estimated fetal weight (EFW) and Last menstrual period gestational age (LMP_GA); Ultrasound mean gestational age (US_MGA). The independent variable (X) is continuous and not fixed. To establish normal reference values for fetal heart parameters, a Z score methodology for the automatic measurement of these parameters was developed.

This study continuously collects data from over 1385 singleton fetuses, with gestational age determined through early pregnancy ultrasound examinations conducted between 18w+0d weeks and 37w+6d weeks. The dataset includes only those cases that provide a complete view of the fetal four-chamber view and excludes any instances of cardiac structural abnormalities, discrepancies in gestational age, or factors that may influence fetal heart rate, such as functional extracardiac and maternal factors. Importantly, this dataset does not overlap with the data used for model building, which includes training, tuning, and validation sets. In cases where multiple four-chamber cardiac videos of the same fetus were obtained at different gestational ages, only the video with the highest image quality was retained. After adjusting gestational age based on early pregnancy ultrasound findings, 95% of independent variables—BPD, HC, AC, FL, and EFW—that exceeded the normal range on both sides were excluded, utilizing criteria published by Hadlock to construct reference ranges for fetal biometric measurements. The criteria for video collection of the four-chamber view remained consistent throughout the study. The steps to calculate Z scores and percentiles are as follows:

Step 1: Model the mean of Y (cardiac parameters) as a function of X (the clinical gestational age, ultrasound mean gestational age, BPD, HC, AC, FL, EFW). After the scatter plot was displayed, fractional polynomials were used to evaluate the specific time reference range that has been applied to fetal bioassays. The NCSS11 software program was used to analyze the sample data automatically. 44 equations arranged in descending order of R^2 values are obtained as show in Table S9.

Step 2: Derive the regression equation of standard deviation by fitting a polynomial of degree, denoted as ($SD = C0 + C1 \times X$).

Step 3: Test the goodness of fit of the model by first considering the values of the higher R^2 or F statistics. Check whether Z-scores are normally distributed by observing the normal probability plot of Z-scores and the results of the normality test.

Step 4: The equation of Z-Score is obtained by using mean regression equation and standard deviation regression equation, denoted as:

$$Z_score = \frac{I_{measure} - M_{normal}}{SD_{normal}}, \quad (6)$$

where, $I_{measure}$ is individual measurement, M_{normal} is predicted mean of the measurement from a normally distributed population, and SD_{normal} is standard deviation of the measurement from a normally distributed population.

Table S10-S12 shows the most regression function of the Z-score modeling for cardiac parameters of full, left and right heart, respectively.

Vedio S1: Demonstration of the process of the automated cardiac function measurement on a normal fetus. The red box highlights key information displayed on the user interface. At the beginning of the video, the user interface displays real-time detection and segmentation results of cardiac structures. After comprehensive evaluation of the detection results, the AI determines that the current video shows a standard four-chamber heart view. Subsequently, by tracking structural changes during several cardiac cycles, it automatically selects the optimal results for cardiac function calculations.

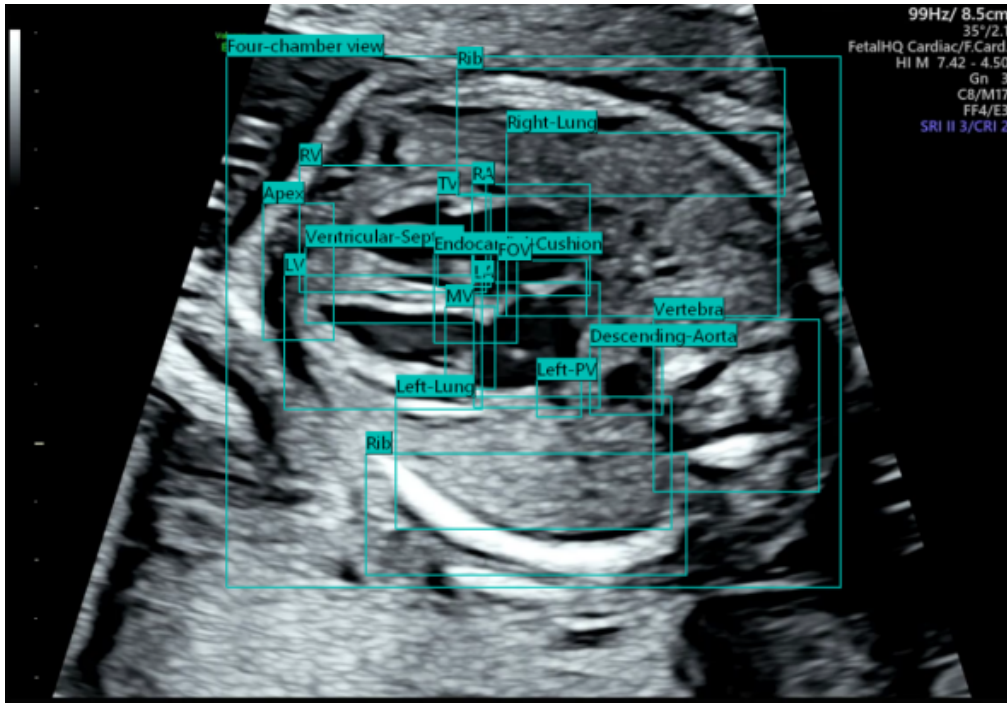


Fig. S1: The detection network's recognition of the anatomical structures

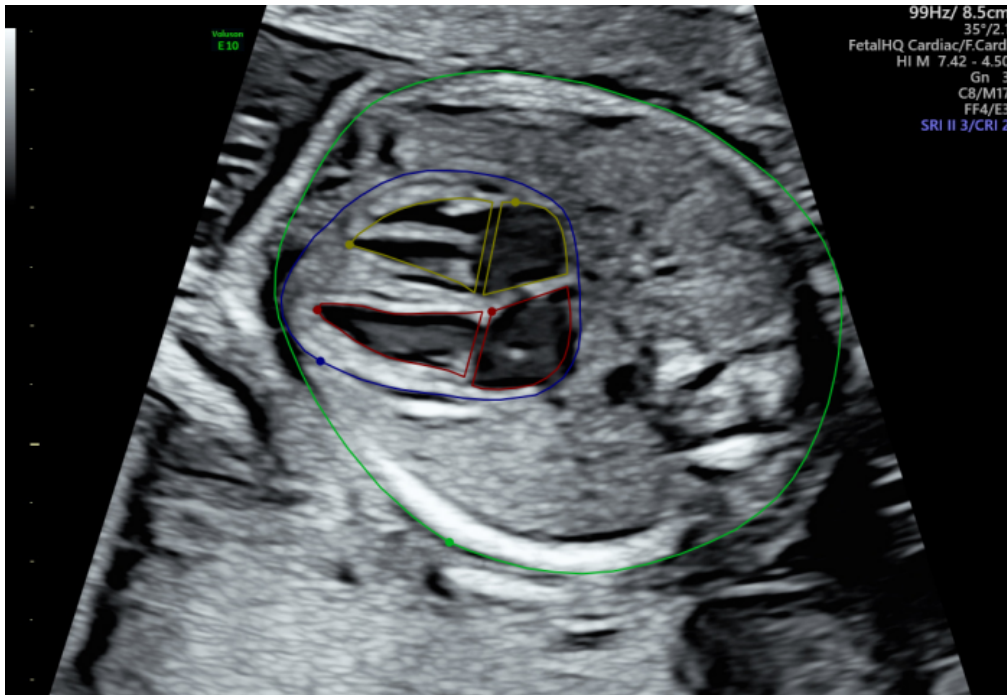


Fig. S2: An segmentation example of the Mask-Dino Head

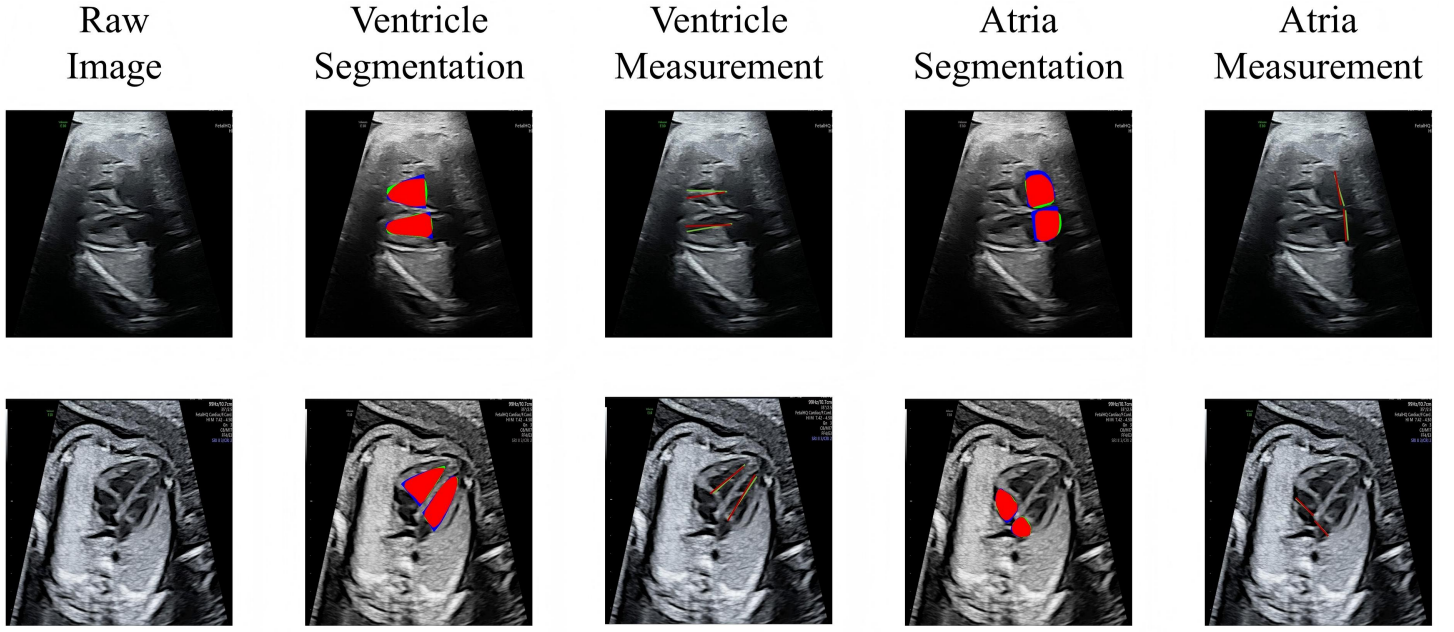


Fig. S3: Schematic diagram of video frame segmentation and measurement results for two echocardiograms. The red/green/blue regions represent correct/over-segmented/under-segmented results, respectively. The red/green straight lines indicate measurement results extracted from manual/predicted masks, respectively.

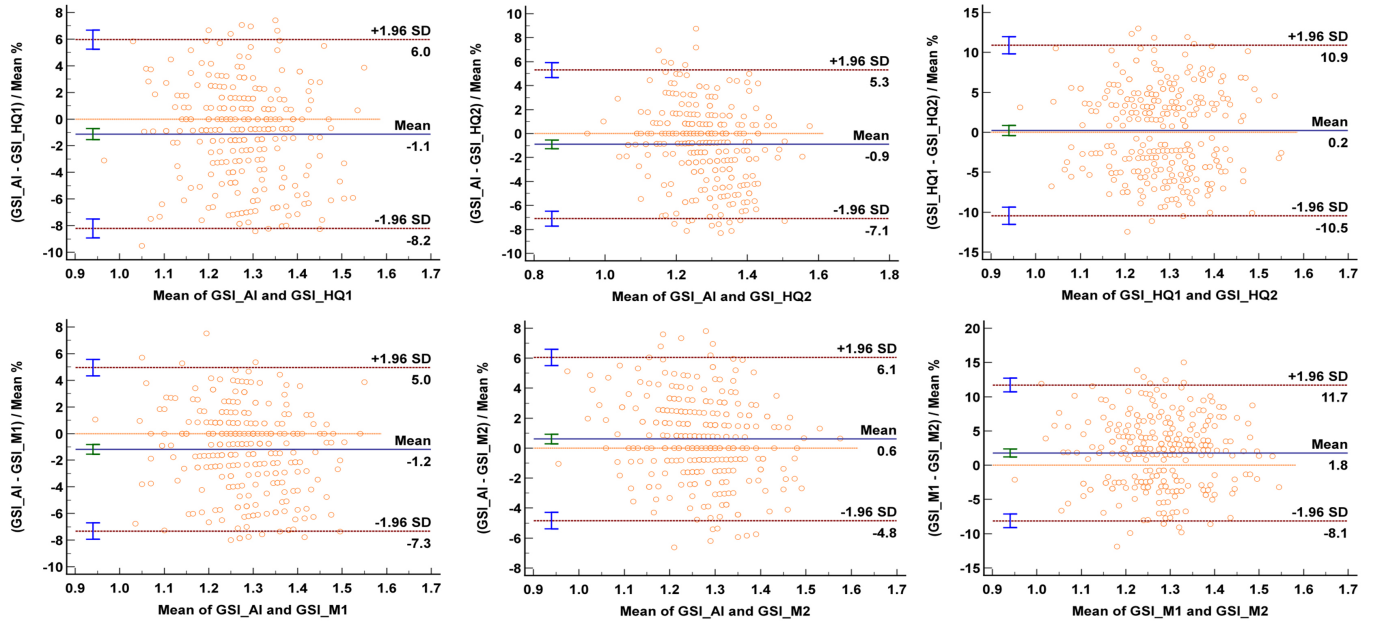


Fig. S4: The Limits of Agreement from Bland-Altman testing of Global Sphericity Index in Internal Normal Dataset

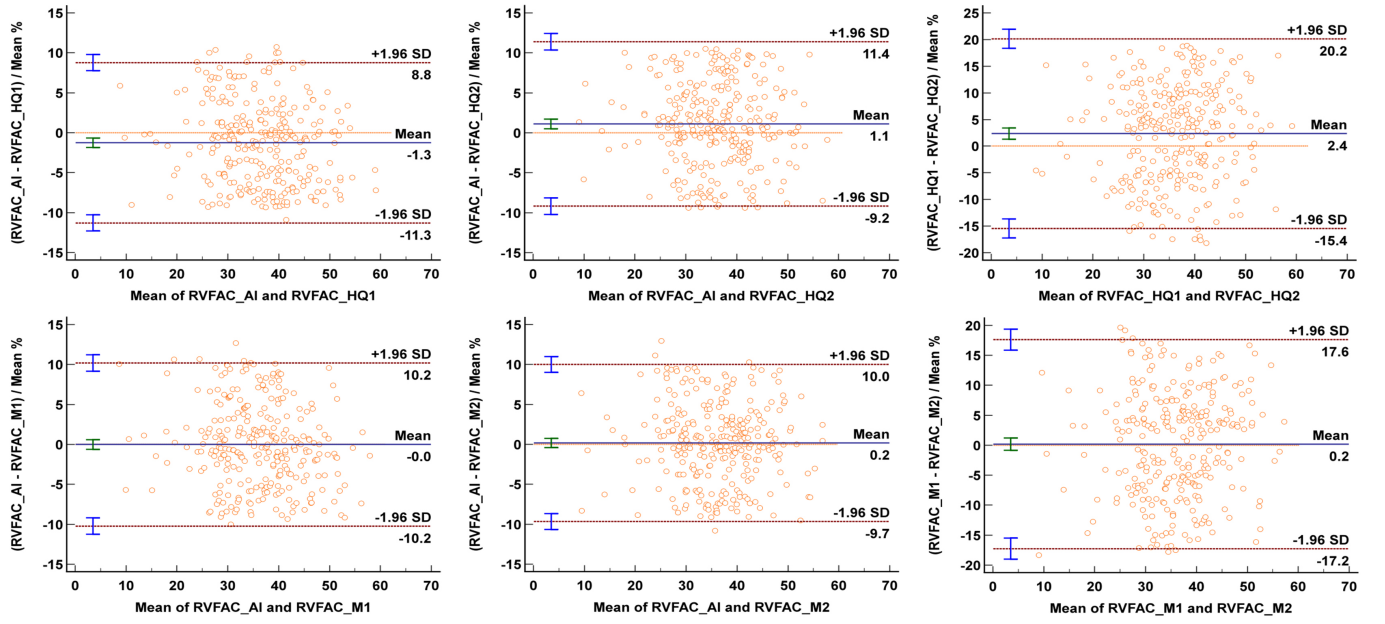


Fig. S5: The Limits of Agreement from Bland-Altman testing of Fractional Area Change of Right Ventricle in Internal Normal Dataset

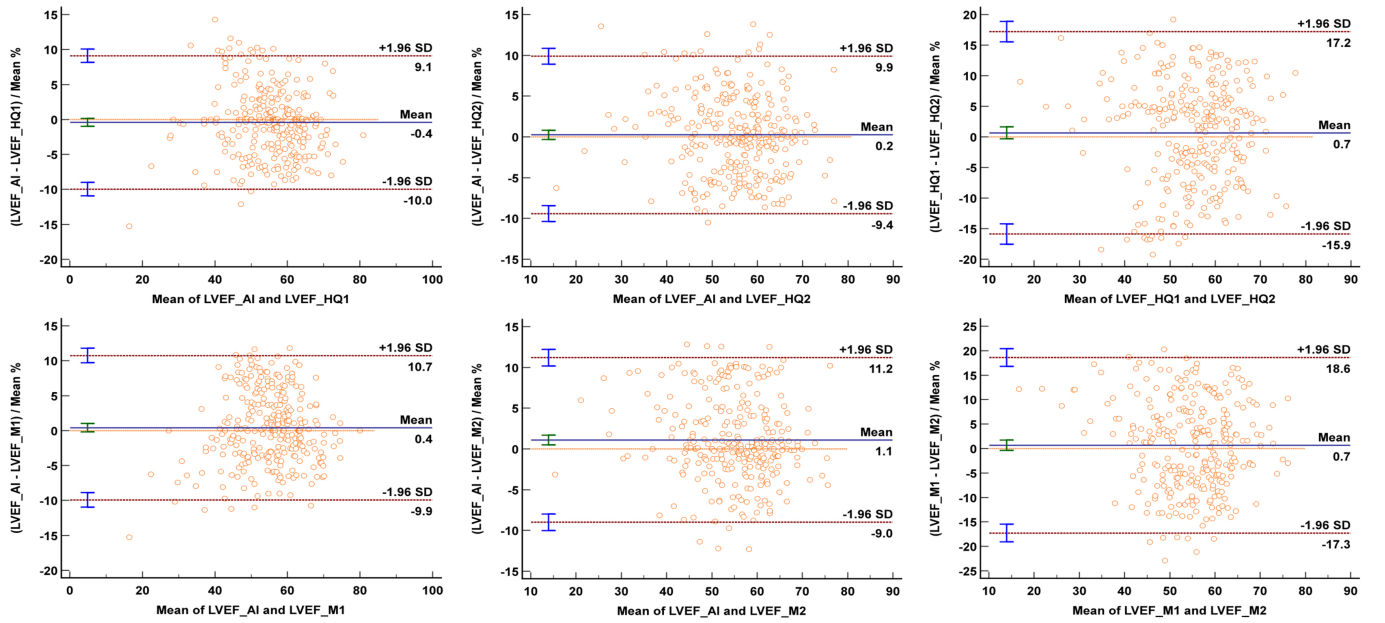


Fig. S6: The Limits of Agreement from Bland-Altman testing of Ejection Fraction of Left Ventricle in Internal Normal Dataset

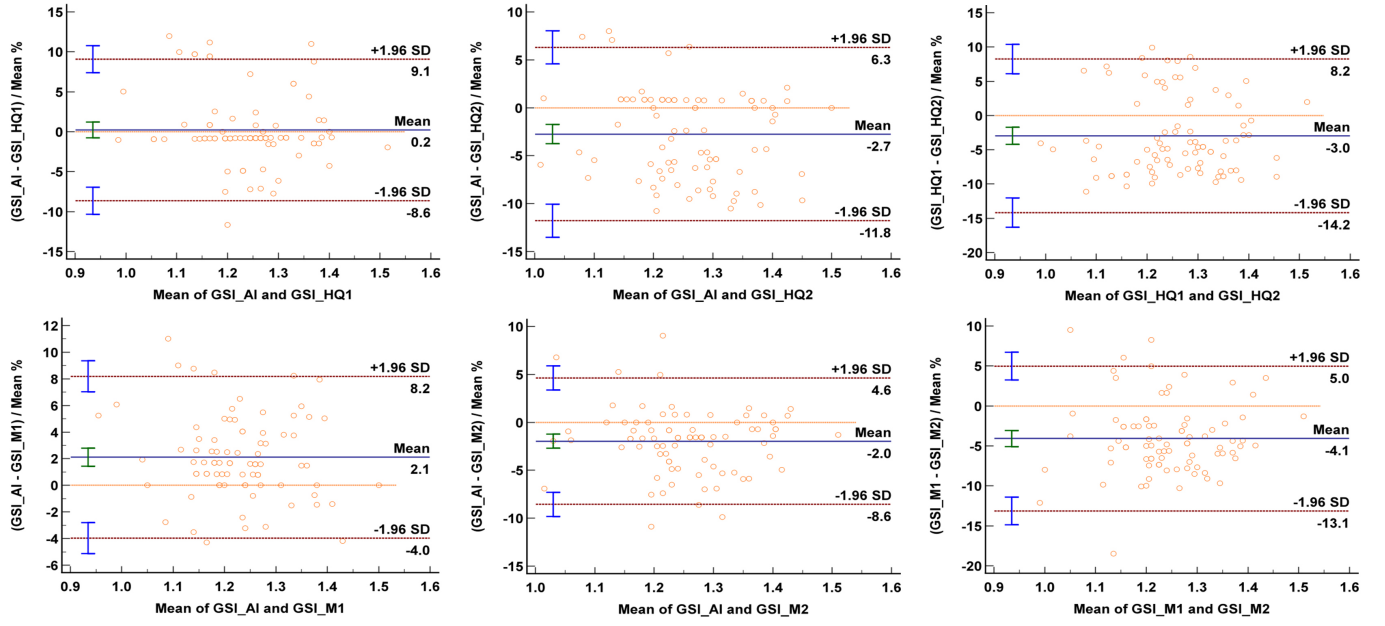


Fig. S7: The Limits of Agreement from Bland-Altman testing of Global Sphericity Index in Internal Abnormal Dataset

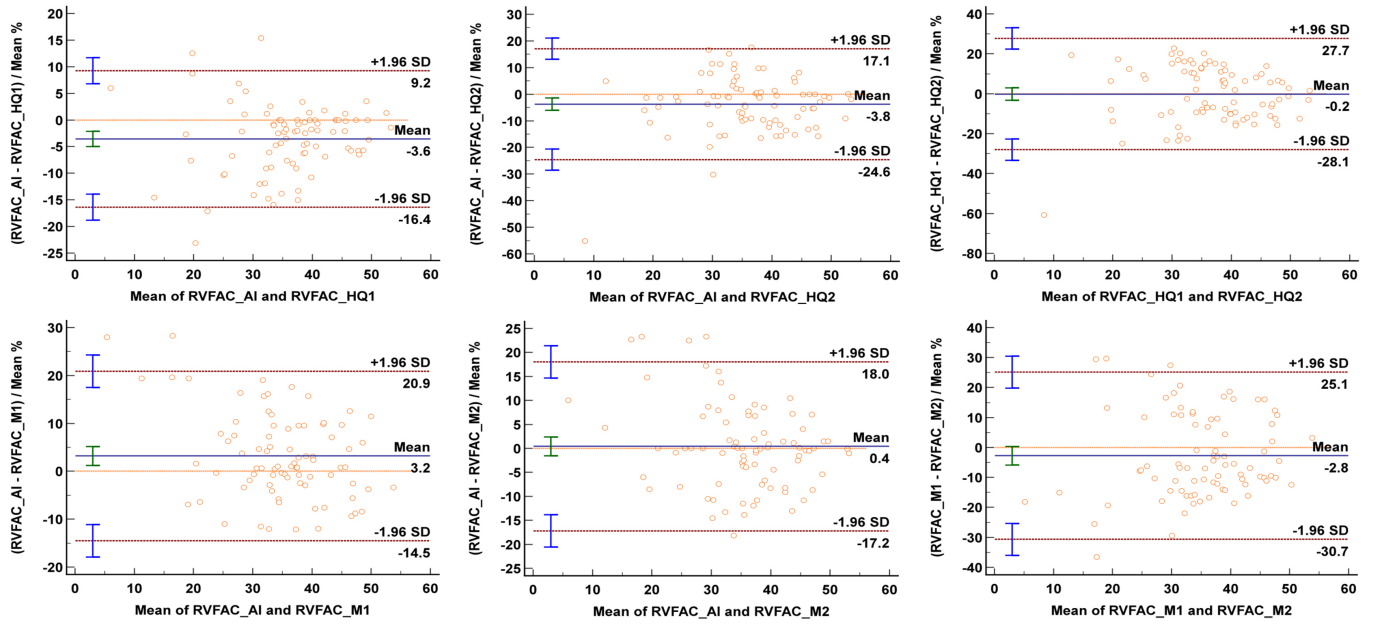


Fig. S8: The Limits of Agreement from Bland-Altman testing of Fractional Area Change of Right Ventricle in Internal Abnormal Dataset

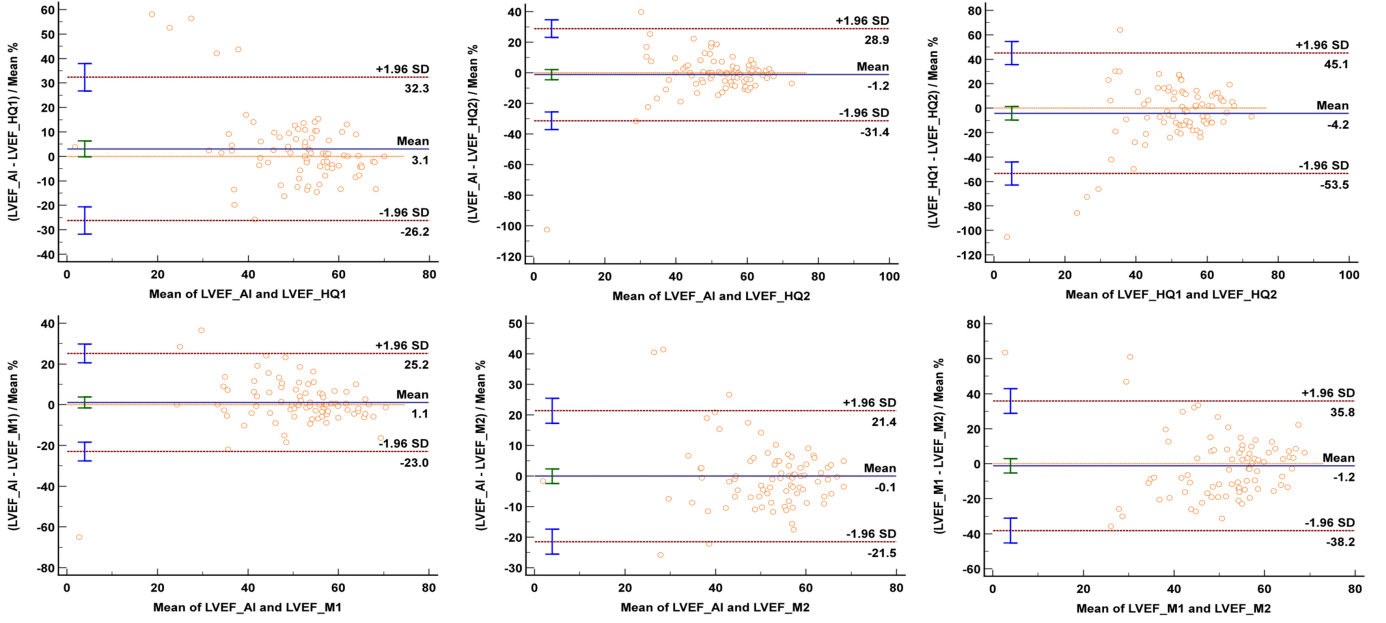


Fig. S9: The Limits of Agreement from Bland-Altman testing of Ejection Fraction of Left Ventricle in Internal Abnormal Dataset

Table S1: Training Hyperparameters and testing device of the AI Model

Hyperparameter	Value	Testing Device	Product name
epoch	200	CPU	Intel Core i7-12700
base learning rate	0.0001	GPU	Nvidia RTX 4070
warmup	true	RAM	16G
warmup steps	2000	Operator System	Window 11
clip grad by norm	0.1		
regularizer	false		
optimizer	AdamW		
weight decay	0.0001		
use focal loss	true		

Table S2: Cardiac parameters obtainable by automatic, manual, and HQ measurement for full heart structures.

Parameter Name		Abbreviation	Heart		
			AI	HQ	Human
Routine Cardiac Parameters	Heart Rate	HR	✓	✓	✓
	Cardiac Axis	CA	✓		✓
	Cardiothoracic Area Ratio	CTAR	✓		✓
	Global Sphericity Index	GSI	✓	✓	✓
	End-diastolic Ventricular Area Ratio	EDVAR	✓	✓	✓
Expand Cardiac parameters	Cardiothoracic Ratio	CTR	✓		✓
	Cardiothoracic Circumference Ratio	CTCR	✓		✓
	End-diastolic Ventricular Width Ratio	EDVWR	✓	✓	✓
	End-diastolic Ventricular Length Ratio	EDVLR	✓	✓	✓
Basic Cardiac parameters	Cardiac Length	CL	✓	✓	✓
	Cardiac Width	CW	✓	✓	✓
	Thoracic Width	TW	✓		✓
	Cardiac Circumference	CC	✓		✓
	Thoracic Circumference	TC	✓		✓
	Cardiac Area	-	✓	✓	✓
	Thoracic Area	TA	✓		✓
	Interventricular Septum Thickness	IVST	✓		✓

Table S3: Cardiac parameters obtainable by automatic, manual, and HQ measurement for left and right heart structures. Ed and Es denotes End-diastolic and End-systolic.

	Parameter Name	Abbreviation	Left Heart			Right Heart		
			AI	HQ	Human	AI	HQ	Human
Routine Cardiac Parameters	Atrial Area	LAA, RAA	✓		✓	✓		✓
	Fractional Area Change	LVFAC, RVFAC	✓	✓	✓	✓	✓	✓
	Stroke Volume	LVSv	✓	✓	✓			
	Ejection Fraction	LVEF	✓	✓	✓			
	Cardiac Output	LVCO	✓	✓	✓			
	Ed Transverse Width	LVTWd, RVTWd	✓	✓	✓	✓	✓	✓
Expand Cardiac parameters	Annular Ventricular Fractional Shortening	LVAVFS, RVAVFS	✓	✓	✓	✓	✓	✓
	Transverse fractional shortening	LVTFS, RVTFS	✓	✓	✓	✓	✓	✓
	Sphericity Index	LVSI, RVSI	✓	✓	✓	✓	✓	✓
	Global Longitudinal Strain	LVGLS, RVGLS	✓	✓	✓	✓	✓	✓
	Septal Wall Strain	LVSWS, RVSWS	✓		✓	✓		✓
	Free Wall Strain	LVFWS, RVFWS	✓		✓	✓		✓
	Annular Plane Systolic Excursion	LVAPSE, RVAPSE	✓		✓	✓		✓
Basic Cardiac parameters	Superior-inferior Diameter	LASID, RASID	✓		✓	✓		✓
	Transverse Diameter	LATD, RATD	✓		✓	✓		✓
	Es Transverse Width	LVTWs, RVTWs	✓		✓	✓		✓
	Ed Basal-apical Length	LVBALd, RVBALd	✓	✓	✓		✓	✓
	Es Basal-apical Length	LVBALs, RVBALs	✓	✓	✓	✓	✓	✓
	Ed Area	LVEAd, RVEAd	✓	✓	✓	✓	✓	✓
	Es Area	LVEAs, RVEAs	✓	✓	✓	✓	✓	✓
	Ed Endocardial Length	LVECLd, RVECLd	✓		✓	✓		✓
	Es Endocardial Length	LVECLs, RVECLs	✓		✓	✓		✓
	Ed Endocardial Septal Wall Length	LVECSWLd, RVECSWLd	✓		✓	✓		✓
	Es Endocardial Septal Wall Length	LVECSWLs, RVECSWLs	✓		✓	✓		✓
	Ed Endocardial Free Wall Length	LVECFWLd, RVECFWLd	✓		✓	✓		✓
	Es Endocardial Free Wall Length	LVECFWLs, RVECFWLs	✓		✓	✓		✓
	Es Free Wall hickness	LVFWd, RVFWd	✓		✓	✓		✓
	Ed Volume	LVEVd	✓	✓	✓			
	Es Volume	LVEVs	✓	✓	✓			

Table S4: Segmentation Performance of Six Key Cardiac Structures on Five State-of-art Methods and Our Model.

Model	Left Atrium				Right Atrium				R-Left Ventricle			
	Dice	IoU	HD	p-values	Dice	IoU	HD	p-values	Dice	IoU	HD	p-values
MaskRCNN	85.08±9.45	75.06±12.67	1.68±1.03	<0.01	88.93±6.18	80.57±9.29	1.38±0.75	<0.01	91.43±5.43	84.63±8.34	1.68±0.98	<0.01
YoloLact	81.56±12.87	73.48±14.34	2.08±0.86	<0.01	84.97±15.13	75.79±17.21	2.54±1.14	<0.01	87.33±15.77	78.70±17.94	2.44±1.13	<0.01
Mask2Former	83.09±8.85	71.94±11.55	1.95±1.17	<0.01	87.49±6.24	78.25±9.04	1.57±0.88	<0.01	89.91±6.32	82.18±8.97	1.89±1.26	<0.01
FastInst	84.51±7.75	73.88±10.61	1.78±0.97	<0.01	87.83±6.25	78.81±9.34	1.50±0.84	<0.01	90.90±4.84	83.65±7.58	1.77±1.07	<0.01
Yolov11	81.82±8.98	70.12±11.84	2.03±1.07	<0.01	86.20±6.45	76.27±9.30	1.72±0.86	<0.01	87.97±7.60	79.25±10.65	2.45±1.66	<0.01
Ours	87.76±6.20	78.70±9.25	1.40±0.71	-	89.99±4.74	82.12±7.48	1.24±0.60	-	92.43±4.17	86.19±6.70	1.44±0.69	-
	Right Ventricle				Cardiac				Thoracic			
	Dice	IoU	HD	p-values	Dice	IoU	HD	p-values	Dice	IoU	HD	p-values
MaskRCNN	90.04±5.76	82.33±8.70	1.91±1.23	<0.01	96.39±1.74	93.09±3.14	1.52±0.80	<0.01	95.45±3.95	91.49±5.55	4.41±3.15	<0.01
YoloLact	84.38±14.55	78.53±16.46	2.69±1.11	<0.01	93.35±4.69	89.72±6.78	3.69±1.02	<0.01	88.92±8.35	85.42±8.95	7.49±2.18	<0.01
Mask2Former	88.66±6.75	80.20±9.50	2.12±1.37	<0.01	95.52±3.27	91.57±4.80	1.88±1.46	<0.01	95.18±2.56	90.92±4.45	4.70±2.87	<0.01
FastInst	88.71±7.20	80.34±9.81	2.14±1.51	<0.01	95.22±3.06	91.02±4.85	1.84±1.31	<0.01	95.36±2.11	91.21±3.76	4.35±2.28	<0.01
Yolov11	86.55±8.91	77.20±11.73	2.61±1.76	<0.01	94.69±3.20	90.08±5.31	2.13±1.50	<0.01	95.43±1.97	91.32±3.53	4.31±2.20	<0.01
Ours	91.28±4.20	84.23±6.81	1.60±0.84	-	96.90±1.47	94.02±2.66	1.34±1.14	-	96.59±2.22	93.50±4.05	3.54±2.63	-

