Supplementary Material

In the Supplementary Material, we first present in Section A detailed descriptions of private CHD dataset and of the EchoNet-Dynamic dataset, including how we converted its ejection-fraction regression labels into a classification task. Section B defines all evaluation metrics (F1, AUC, ECE, AECE). Section C covers implementation details such as frame sampling, optimizer settings, hyperparameters, and prompt templates. Finally, Section D reports the complete ablation results. Our code will be released publicly upon publication.

A Datasets

Private Dataset. As no public benchmark for fetal CHD detection exists, a private dataset of 477 echocardiographic videos was curated. Each video, showing the four-chamber (4CH) view of the fetal heart, was acquired at a local hospital and annotated by a senior Maternal–Fetal Medicine consultant. Videos were labeled as either Normal (198 cases) or as one of 19 distinct congenital heart disease (CHD) classes. Among these, the most frequently represented defects are ventricular septal defect (VSD) (43 cases), atrioventricular septal defect (AVSD) (33), arrhythmia (33), and cardiomegaly (31). Model performance was assessed via 5-fold cross-validation with class-stratified splits.

EchoNet-Dynamic Dataset. The EchoNet-Dynamic dataset [31], comprising 10,030 echocardiographic videos annotated with estimated ejection fraction (EF), was employed for further evaluation. Following the PanEcho protocol [16], EF values were binned into three systolic dysfunction categories: "Moderately/Severely Decreased" ($EF < 40\%$), "Mildly Decreased" ($40\% \leq EF \leq 54\%$), and "Normal/Hyperdynamic" ($EF > 54\%$).

B Metrics

Discrimination was assessed using the macro F1 score, which balances precision and recall, and the area under the receiver operating characteristic curve (AUC), which quantifies the ability to rank positive instances above negative ones across all thresholds. Calibration was measured by the expected calibration error (ECE), defined as the average absolute difference between predicted confidence and empirical accuracy over fixed bins, and by the adaptive expected calibration error (AECE), which employs quantile-based bins to prevent underpopulated intervals.

F1 Score The harmonic mean of precision and recall,

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{8}$$

which balances false positives and false negatives.

Macro F1 Score The unweighted average of per-class F1 scores,

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{F1}_c \tag{9}$$

giving equal weight to each class in multiclass settings.

Area under the ROC Curve (AUC) The probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative one. Values range from 0.5 (random) to 1.0 (perfect). For multiclass tasks, we use the macro-average of one-vs-rest AUCs.

Expected Calibration Error (ECE) [13]

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{10}$$

where B_m is the set of predictions in bin m, n is the total number of samples, and $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ are its accuracy and average confidence. In our experiments we set $M = 15$ uniform-width bins.

Adaptive Expected Calibration Error (AECE) [30] Computed with quantile-based bins so each bin has roughly the same number of samples.

C Implementation Details

Implementation Details. Following [15, 16, 28], $L = 16$ consecutive frames were sampled from each video. A batch size of 16 was used, the Adam optimizer (initial learning rate 1×10^{-3}) was employed, and a ReduceLROnPlateau scheduler (decay factor 0.1, patience 5) was applied. On the private CHD dataset, training was carried out for up to 40 epochs with early stopping based on validation macro F1, for EchoNet-Dynamic, the same hyperparameters were retained, but training was restricted to 20 epochs. All experiments were conducted on an NVIDIA RTX 5000 Ada Generation GPU. For uncertainty quantification, β was set to 0.2 in Equation 7.

Binary CHD detection prompts. For binary CHD detection we used, "Is the fetal heart normal in this 4CH ultrasound view?" and "Can an abnormality be detected in the fetal heart 4CH?" For prompt randomization, we sample one prompt per class from six variants:

Normal: "Is the fetal heart normal in this 4CH ultrasound view?"; "Does this 4CH image show a normal heart condition?"; "Is everything normal in this fetal heart 4CH scan?"; "Does this ultrasound suggest a healthy fetal heart?"; "Does this 4CH scan reflect a structurally normal heart?"; "normal"

Abnormal: "Is abnormality visible in this fetal heart 4CH scan?"; "Does this image show abnormal fetal heart?"; "Can abnormality be detected in this fetal heart 4CH?"; "Does this 4CH scan suggest the presence of abnormality?"; "Does this ultrasound reveal abnormality in the fetal heart?"; "abnormal"

Multiclass prompts. Each defect class is associated with one fixed, class-specific description:

VSD A defect in the interventricular septum permits blood to flow from the left to the right ventricle, causing abnormal left-to-right shunting and right-heart volume overload.

AVSD This anomaly features incomplete fusion of the atrioventricular septum and malformed AV valves, resulting in a common AV junction and mixed atrial-ventricular flow.

Arrhythmia Rhythmic irregularities demonstrate arrhythmia, from isolated premature beats to sustained tachy- or bradyarrhythmic patterns.

Cardiomegaly An enlarged cardiac silhouette occupying over half the thoracic area indicates cardiomegaly, often due to volume overload, cardiomyopathy, or high-output states.

EchoNet-Dynamic prompts. We use three fixed descriptions that align with the ejection-fraction categories:

- "Normal/Hyperdynamic"
- "Mildly Decreased"
- "Moderately/Severely Decreased"

D Ablation

Hyperparameter	F1(↑)	AUC(↑)
$m = 0.5$	**84.69**	86.14
$m = 1.0$	84.45	**86.88**
$\alpha = 1$	84.20	86.89
$\alpha = 0.5$	**84.69**	**87.59**
$\alpha = 0$	84.67	86.11
random prompts	84.24	86.96

(a) m, α, and prompt randomization.

Hyperparameter	F1(↑)	AUC(↑)
4 neighbors, mean pooling	83.42	86.55
16 neighbors, mean pooling	83.71	86.92
10 neighbors, mean pooling	84.42	**89.02**
10 neighbors, max pooling	**85.01**	87.49
3 passes through GNN	84.31	87.71
graph fusion by summation	84.58	88.46

(b)Ablation on GNN components.

Hyperparameter	F1(↑)	AUC(↑)
mLSTM	84.98	87.70
sLSTM	**85.43**	**88.33**
mid fusion	84.82	86.31
late fusion	85.15	88.06
2-block	84.46	85.69
7-block	85.40	88.31
12-block	85.35	86.10

(c) Ablation on xLSTM components.

Temporal Extractor	Private dataset		EchoNet-Dynamic	
	F1	AUC	F1	AUC
1D-CNN	81.56	82.44	56.68	81.38
TCN	81.83	85.44	55.98	80.51
GNN	**84.84**	**87.69**	**58.62**	**82.11**
xLSTM	80.32	84.64	54.01	79.51

(d) Ablation on Temporal Extractors within TPA.

Table 5: Ablation study of TPA performance: (a) Loss hyperparameters: m, α in Equation 2, and prompt randomization; (b) GNN design choices: number of neighbors, pooling method and fusion strategy; (c) xLSTM variants: gating type (mLSTM vs sLSTM), fusion stage and network depth; (d) Comparison of different temporal extractors within TPA on the private and EchoNet-Dynamic datasets. Best results are highlighted in **bold**.

Class	**TCN**		**GNN**		**xLSTM**	
	F1 (↑)	AUC (↑)	F1 (↑)	AUC (↑)	F1 (↑)	AUC (↑)
Defects = 2						
Macro	**78.68 ± 8.20**	71.09 ± 11.35	76.46 ± 9.92	72.76 ± 13.72	76.82 ± 8.73	72.55 ± 14.69
Normal	**88.69 ± 5.24**	76.36 ± 9.69	88.09 ± 6.55	78.25 ± 11.76	88.34 ± 6.54	79.22 ± 13.07
VSD	**53.96 ± 1.13**	62.93 ± 1.63	48.78 ± 17.42	65.06 ± 20.30	51.69 ± 1.54	65.92 ± 1.96
AVSD	**59.56 ± 2.06**	73.98 ± 14.78	59.56 ± 18.33	74.96 ± 14.79	57.80 ± 1.21	67.95 ± 1.55
Defects = 3						
Macro	**72.91 ± 3.75**	70.07 ± 8.94	71.80 ± 4.60	69.53 ± 8.13	71.27 ± 6.03	69.37 ± 9.94
Normal	84.62 ± 2.47	83.79 ± 9.54	85.08 ± 2.92	79.50 ± 7.11	**85.22 ± 4.81**	77.44 ± 8.41
VSD	**56.86 ± 15.34**	70.01 ± 19.50	51.14 ± 10.78	69.97 ± 20.50	54.50 ± 1.33	73.24 ± 2.02
AVSD	**59.26 ± 16.84**	69.83 ± 17.06	56.03 ± 8.92	68.56 ± 15.47	56.16 ± 1.21	67.95 ± 1.15
Arrhythmia	43.99 ± 11.13	56.64 ± 10.69	44.17 ± 15.02	60.09 ± 12.77	**45.78 ± 1.34**	58.85 ± 1.32
Defects = 4						
Macro	**67.63 ± 5.33**	66.20 ± 5.46	67.49 ± 5.18	66.77 ± 2.71	66.90 ± 4.64	51.61 ± 10.63
Normal	84.72 ± 2.77	86.61 ± 5.43	**86.61 ± 3.23**	86.48 ± 6.58	84.49 ± 4.04	86.64 ± 4.99
VSD	**46.70 ± 11.88**	57.70 ± 17.77	46.61 ± 11.59	60.33 ± 14.20	43.27 ± 8.60	58.23 ± 19.00
AVSD	**55.98 ± 17.38**	68.19 ± 16.30	51.94 ± 7.93	66.27 ± 12.18	46.32 ± 18.12	62.47 ± 15.51
Arrhythmia	31.14 ± 9.88	42.80 ± 12.87	34.21 ± 7.36	47.24 ± 12.22	**37.53 ± 9.58**	49.27 ± 9.44
Cardiomegaly	55.68 ± 17.09	75.72 ± 16.74	54.77 ± 10.63	73.52 ± 9.09	**58.68 ± 14.38**	69.59 ± 1.38

Table 6: Ablation of temporal feature extractors in TPA on the private CHD dataset for multiclass classification with 2, 3, and 4 defects plus the normal class, reporting macro-F1 and AUC across 5 folds. Best results are highlighted in **bold**.