CardiacPULSE: Frequency-Aware Self-Supervised Cardiac Phase Detection in Echocardiography

Jiarong Guo[1 [0000−0002−0231−9096]], Jiewen Yang[1 [0009−0004−1500−1556]], Yao Du[1 [0000−0001−7137−8172]], Yi Qin[1 [0009−0000−2236−652X]], and Xiaomeng Li[1(✉) [0000−0003−1105−8083]]

The Hong Kong University of Science and Technology, Hong Kong SAR, China
eexmli@ust.hk

Abstract. In quantitative echocardiography, detecting end-diastole (ED) and end-systole (ES) frames is crucial for downstream clinical tasks. However, ECG synchronization and frame-level annotations are often unavailable in portable ultrasound devices. While recent self-supervised methods extract cardiac phase cues from latent motion trajectories, they suffer from view-specific reconstruction objectives, leading to unstable estimates across different views. To address this, we propose **CardiacPULSE**, a **Cardiac P**hase detection framework that achieves **U**ltralight inference by **L**ocalizing **S**ignal **E**xtrema from the inherent periodicity of cardiac motion. Our method constructs a physics-guided spatial prior via per-pixel temporal Fourier analysis to isolate rhythmically active regions. To enable efficient edge deployment, we train a lightweight auxiliary network to predict this Fourier-derived motion mask in a single forward pass under purely self-supervised losses. Our approach achieves a 34.0 ms mean absolute error on the EchoNet-Dynamic dataset, outperforming CardiacPhase by 31%. The code is available at https://github.com/xmed-lab/CardiacPulse.

Keywords: Echocardiography · Cardiac Phase Detection · Self-supervised Learning · Fourier Analysis.

1 Introduction

Accurate identification of end-diastole (ED) and end-systole (ES) frames is fundamental to quantitative echocardiography, as these temporal landmarks enable the computation of key clinical metrics including left-ventricular ejection fraction, stroke volume, and cardiac output—all critical for diagnosing heart failure, valvular disease, and cardiomyopathy [15,18,3,14,16,9,4]. In standard clinical workflows, ED and ES are typically anchored by synchronized electrocardiogram (ECG) signals [24,14]. However, ECG integration remains impractical for portable ultrasound devices due to hardware constraints and workflow complexity [2]. Consequently, phase timing must be inferring phase timing directly from echocardiography videos presents many challenges: inferred directly from echocardiographic video, which poses significant challenges, *e.g.*, speckle noise

obscures myocardial boundaries, view-dependent appearance varies across imaging planes, and probe motion introduces non-cardiac artifacts [8,13,21]. Furthermore, deploying such algorithms on edge devices requires lightweight architectures that balance accuracy with computational efficiency, which is often overlooked [22].

Early ECG-free approaches either relied on labor-intensive pixel-level segmentation to track cavity volumes [12,23,17,7] or required manual frame-level phase labels for supervised classification [7,10]. While classical speckle-tracking methods [25] avoid deep learning, they depend on view-specific heuristics. Recently, learning-based formulations have explored annotation-free phase detection by modeling cardiac motion in latent space. CardiacPhase [20] introduced a self-supervised paradigm that infers ED/ES as geometric extrema in a learned latent motion space, eliminating the need for structural or temporal annotations. However, this reconstruction-based strategy inherently couples motion representations to specific imaging views (e.g., apical four-chamber vs. parasternal), severely limiting generalization across varying acquisition protocols without retraining [19,20,6]. Furthermore, the reliance on complex latent trajectory extraction and multi-stage post-processing imposes significant computational overhead, hindering deployment on resource-constrained edge devices.

To address these limitations, we propose **CardiacPULSE**, a novel framework that revisits unsupervised cardiac phase detection from a frequency-aware and deployment-oriented perspective. Rather than inferring ED/ES indirectly from geometric extrema in a learned latent motion space, where periodic structure arises only as a byproduct of view-specific reconstruction, we explicitly encode dynamics periodic at the cardiac frequency as an inductive bias at the signal level. This foundation stems from two complementary innovations: First, myocardial motion exhibits strong spectral concentration at the dominant cardiac frequency, enabling per-pixel temporal Fourier analysis to derive a physics-guided spatial prior that isolates rhythmically active regions without segmentation or labels. Second, to enable real-world deployment on portable ultrasound devices, we distill this spectral prior into a lightweight auxiliary network (103K parameters) trained under purely self-supervised objectives (spectral concentration, phase coherence, and mask coverage), making it highly suitable for real-time inference on resource-constrained portable ultrasound devices. Critically, this design decouples expensive offline Fourier computation from online inference, achieving real-time performance with minimal memory footprint.

Finally, we construct analytically grounded temporal cues by fusing a high-recall, phase-agnostic motion-energy candidate pool with phase-specific anchors obtained via low-order harmonic reconstruction of a mask-weighted intensity signal. This additive fusion preserves recall while resolving systole–diastole ambiguity caused by asymmetric cardiac motion. Critically, CardiacPULSE requires no manual annotations and achieves precise frame-level detection through interpretable, frequency-domain principles, which makes it uniquely suited for scalable, edge-compatible quantitative echocardiography. We achieve 34.0 ms average MAE (ED: 34.4 ms, ES: 33.6 ms) on EchoNet-Dynamic [15], improving

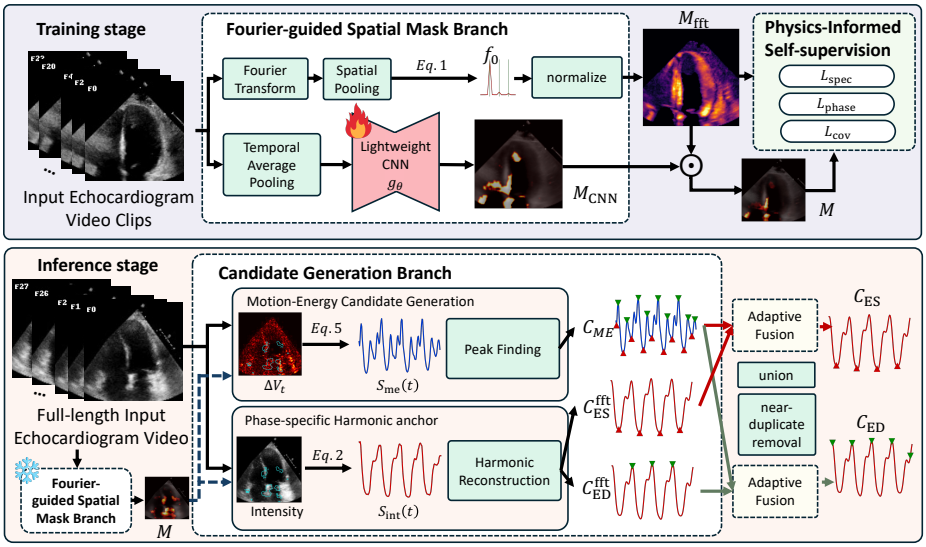

Fig. 1: **The Overall pipeline of CardiacPULSE**. Top: Fourier-guided spatial mask during training, mask are shown as overlaying on echocardiogram frames. Bottom: inference stage, fuses phase-agnostic motion-energy extrema with phase-specific harmonic anchors to detect ED/ES frames.

over the self-supervised CardiacPhase baseline (49.0 ms) by 31% while using only 103K trainable parameters.

2 Methodology

We address ECG-free ED/ES detection in echocardiography without phase labels or segmentation annotations. Given a video $\mathbf{X} \in \mathbb{R}^{T\times H\times W}$, our framework (Fig. 1) separates training and inference: we first learn, in a self-supervised manner, a spatial weighting mask $M \in [0,1]^{H\times W}$ that highlights regions periodic at the cardiac frequency, then derive ED/ES candidates analytically from mask-weighted temporal signals. The only learnable component is a lightweight CNN.

2.1 Fourier-Guided Spatial Mask Learning

Myocardial tissue undergoes periodic deformation synchronized with the cardiac cycle, resulting in strong spectral energy at the dominant heart rate frequency. In contrast, static structures or probe-induced motion lack such rhythmicity. We leverage this physical property to construct a data-driven spatial prior that isolates cardio-synchronous regions without supervision.

For each spatial location (x, y), we compute the discrete Fourier transform (DFT) along the temporal dimension of $\mathbf{X}$, denoted $\mathcal{F}_{x,y}(f)$, where f indexes frequency bins in Hz. Within a physiologically plausible band $\mathcal{B} = [0.5, 4.0]$ Hz, we

estimate the global dominant cardiac frequency f_0 as the frequency maximizing the spatially averaged power spectrum:

$$f_0 = \text{argmax}_{f\in\mathcal{B}} \frac{1}{HW} \sum_{x=1}^{H} \sum_{y=1}^{W} |\mathcal{F}_{x,y}(f)|^2. \tag{1}$$

The magnitude response at f_0 yields a physics-based spatial prior M_{fft}, normalized to $[0,1]$ as $M_{\text{fft}}(x,y) = \frac{|\mathcal{F}_{x,y}(f_0)|-\min|\mathcal{F}|}{\max|\mathcal{F}|-\min|\mathcal{F}|}$. This map emphasizes pixels whose intensity oscillates coherently at the estimated heart rate.

While M_{fft} captures motion rhythm, it may be corrupted by noise or non-cardiac oscillations. To refine it with anatomical plausibility, we introduce a small fully convolutional network g_θ that predicts an appearance-consistent mask $M_{\text{CNN}} = g_\theta(\bar{\mathbf{X}})$, where $\bar{\mathbf{X}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}(t,\cdot,\cdot) \in \mathbb{R}^{H\times W}$ is the temporally averaged frame. The final spatial mask is obtained by element-wise multiplication, $M = M_{\text{CNN}} \odot M_{\text{fft}}$, which ensures that only regions both rhythmically active and anatomically plausible contribute to downstream signals. A small constant floor (e.g., 10^{-3}) is applied to M_{CNN} during training to prevent collapse to zero. We then use M to construct a scalar temporal signal by spatially averaging the input video:

$$s_{\text{int}}(t) = \frac{\sum_{x=1}^{H}\sum_{y=1}^{W} M(x,y)\mathbf{X}(t,x,y)}{\sum_{x=1}^{H}\sum_{y=1}^{W} M(x,y) + \varepsilon}, \tag{2}$$

where $\varepsilon > 0$ is a numerical stabilizer. This signal reflects the global intensity dynamics of the masked cardiac region over time.

2.2 Physics-Informed Self-Supervision from Cardiac Periodicity

All objectives are applied to signals extracted using the combined mask M, so optimizing g_θ effectively learns M_{CNN} to produce an M that best exposes cardiac periodicity relevant to ED/ES. Since no ground-truth masks or phase labels are available, we design three self-supervised objectives that enforce spectral purity, temporal coherence, and spatial coverage—properties inherent to true cardiac motion. Let $S(f) = \mathcal{F}s_{\text{int}}(t)$ denote the DFT of s_{int}, and $P(f) = |S(f)|^2$ its power spectrum. We define the dominant frequency in the learned signal as $f^\star = \text{argmax}_{f\in\mathcal{B}} P(f)$. First, the spectral concentration loss $\mathcal{L}_{\text{spec}} = -\log(\frac{P(f^\star)}{\sum f\in\mathcal{B}P(f)+\varepsilon})$ encourages energy to concentrate at a single frequency, mimicking the quasi-periodic nature of cardiac cycles. Second, the phase coherence loss promotes spatial alignment of oscillation phases across the mask. Let $\Phi(f_0;x,y) = \angle\mathcal{F}_{x,y}(f_0)$ denote the phase of the raw pixel signal at f_0. We compute the circular mean of these phases, weighted by M_{CNN}, and penalize dispersion:

$$\mathcal{L}_{\text{phase}} = 1 - \left| \frac{\sum_{x,y} M_{\text{CNN}}(x,y) e^{j\Phi(f_0;x,y)}}{\sum_{x,y} M_{\text{CNN}}(x,y) + \varepsilon} \right|. \tag{3}$$

A value close to 0 indicates high phase alignment, as expected in coordinated myocardial motion. Third, to avoid degenerate solutions (e.g., empty masks), we

impose a coverage loss that balances mask activation around a target mean τ in $(0, 1)$ while preventing collapse via a log-barrier term:

$$\mathcal{L}_{\text{cov}} = (\bar{m} - \tau)^2 - \beta \log(\bar{m} + \varepsilon), \quad \text{where} \quad \bar{m} = \frac{1}{HW} \sum\nolimits_{x,y} M_{\text{CNN}}(x, y), \quad (4)$$

where $\beta > 0$ controls the strength of the barrier. The total loss is a weighted sum $\mathcal{L} = \lambda_{\text{spec}}\mathcal{L}_{\text{spec}} + \lambda_{\text{phase}}\mathcal{L}_{\text{phase}} + \lambda_{\text{cov}}\mathcal{L}_{\text{cov}}$, with hyperparameters specified.

2.3 Unsupervised Candidate Generation via Motion Dynamics

Direct extrema of $s_{\text{int}}(t)$ are unreliable due to systole–diastole asymmetry or baseline drift. To ensure high recall of potential ED/ES frames, we construct a motion-energy signal that captures local temporal changes, which are more robust to intensity bias. During inference, we compute a motion-energy waveform $s_{\text{me}}(t)$ as the mask-weighted average of absolute frame-to-frame differences:

$$s_{\text{me}}(t) = \frac{\sum_{x,y} M(x, y)|\mathbf{X}(t + 1, x, y) - \mathbf{X}(t, x, y)|}{\sum_{x,y} M(x, y) + \varepsilon}, \quad t = 1, \ldots, T - 1. \quad (5)$$

This serves as a proxy for instantaneous velocity magnitude. After applying Savitzky–Golay smoothing and linear detrending to remove slow drift, we detect all local maxima and minima of s_{me}. These turning points form a phase-agnostic candidate set C_{ME}, shifted by $+0.5$ frame to account for the midpoint of differencing intervals.

While C_{ME} provides high recall, it does not distinguish ED from ES. To resolve this ambiguity, we exploit the fact that true cardiac motion is approximately sinusoidal and can be well approximated by its first few harmonics. Reconstructing a clean periodic signal allows us to assign phase-specific roles to peaks (ED) and troughs (ES). We perform an inverse DFT on $s_{\text{int}}(t)$ after retaining only the lowest three harmonics $f_0, 2f_0, 3f_0$, yielding a denoised, periodic reconstruction $\tilde{s}_{\text{int}}(t)$. Peaks of $\tilde{s}_{\text{int}}$ correspond to end-diastole, and troughs to end-systole. These define phase-specific anchor sets $C_{\text{ED}}^{\text{fft}}$ and $C_{\text{ES}}^{\text{fft}}$.

Finally, the motion-energy turning points and the phase-specific harmonic anchors are combined into the final ED and ES candidate sets, which retain broad temporal coverage while resolving the ED–ES assignment through their complementary cues.

3 Experiments

We evaluate the performance of our model on two echocardiogram video datasets: **(1) EchoNet-Dynamic**, a public benchmark containing 10,030 apical four-chamber echocardiography videos with expert-derived ED and ES frames [15] in one cycle. We follow the official split and report results on the held-out test set.

[1] Results follow the CardiacPhase [20] that calculates the Mean Absolute Error (MAE) between the ground truth and the temporally closest prediction [20]. We report the MAE in milliseconds (ms) and frames with standard deviation (std).

Table 1: Cardiac phase detection results[1] on EchoNet-Dynamic [15].

Method	Supervision	MAE (frames)		MAE (ms)		Avg.	Params
		ED	ES	ED	ES	MAE (ms)	
R3D-50† [12]	Yes	2.1 (2.5)	1.7 (2.7)	40.2 (48.9)	32.6 (51.6)	36.4	47 M
MAEF† [23]	Yes	2.3 (2.3)	2.4 (2.2)	44.5 (44.2)	46.1 (42.3)	45.3	40 M
UVT† [17]	Yes	7.2 (12.9)	3.4 (6.8)	139.2 (250.4)	65.0 (131.9)	102.1	350.7 M
RepNet-Row [5]	No	5.0 (8.2)	7.0 (7.7)	97.1 (156.2)	135.5 (144.7)	116.3	16 M
RepNet-Lag [5]	No	6.1 (11.5)	5.7 (5.9)	119.8 (226.1)	110.3 (112.6)	115.1	16 M
DDSB§ [1]	No	4.3 (4.8)	8.9 (9.5)	83.4 (93.3)	175.4 (185.4)	129.4	–
CardiacPhase† [20]	No	3.0 (3.3)	2.0 (2.2)	58.3 (66.4)	39.8 (43.0)	49.1	8.6 M
CardiacPULSE (ours)	**No**	**1.76 (1.71)**	**1.72 (1.46)**	**34.4 (34.4)**	**33.6 (28.5)**	**34.0**	**0.1 M**

†Cited from CardiacPhase [20]. §Cited from DDSB [1].

Table 2: Multi-view cardiac phase detection results[1] on CAMEO [11].

Group	View	Method	MAE (frames)		MAE (ms)		Avg.
			ED	ES	ED	ES	MAE (ms)
Apical	A2C	**Ours**	**0.95 (0.64)**	**1.35 (1.25)**	**17.8 (11.6)**	**26.1 (23.9)**	**22.0**
		CardiacPhase [20]	3.27 (3.58)	2.91 (3.05)	143.3 (159.7)	126.2 (134.9)	134.8
	A3C	**Ours**	**1.72 (1.46)**	**1.78 (1.62)**	**32.6 (27.9)**	**35.1 (35.7)**	**33.8**
		CardiacPhase [20]	4.09 (4.68)	3.27 (3.44)	158.7 (190.7)	121.0 (124.4)	139.9
	A4C	**Ours**	**2.20 (1.40)**	**1.90 (1.22)**	**42.5 (28.9)**	**35.6 (22.6)**	**39.0**
		CardiacPhase [20]	4.64 (6.02)	2.45 (4.01)	194.9 (266.9)	90.1 (139.8)	142.5
	S4C	**Ours**	**3.39 (2.93)**	**2.72 (1.92)**	**61.8 (52.2)**	**51.3 (37.8)**	**56.6**
		CardiacPhase [20]	5.00 (4.58)	5.27 (5.59)	195.1 (160.6)	210.3 (204.8)	202.7
Long-axis	PLAX	**Ours**	**1.60 (1.15)**	**2.20 (1.18)**	**30.3 (21.7)**	**41.7 (22.0)**	**36.0**
		CardiacPhase [20]	3.00 (4.29)	2.50 (3.50)	114.0 (163.0)	93.8 (131.3)	103.9
Short-axis (PSAX)	Apical	**Ours**	**3.15 (2.55)**	**2.55 (2.06)**	**59.1 (47.3)**	**48.3 (40.7)**	**53.7**
		CardiacPhase [20]	3.70 (3.13)	3.10 (3.48)	149.7 (132.3)	114.9 (121.6)	132.3
	Aortic	**Ours**	**2.17 (2.12)**	2.56 (1.63)	**41.0 (42.3)**	**49.2 (34.1)**	**45.1**
		CardiacPhase [20]	3.90 (4.65)	**1.90 (3.11)**	148.9 (171.5)	68.7 (109.0)	108.8
	Mitral	**Ours**	**2.75 (1.51)**	**1.65 (1.94)**	**51.3 (27.8)**	**33.9 (44.6)**	**42.6**
		CardiacPhase [20]	4.91 (5.01)	3.18 (4.17)	202.8 (209.5)	133.6 (171.6)	168.2
	Papillary	**Ours**	**2.72 (2.11)**	**1.72 (1.30)**	**50.1 (37.8)**	**33.9 (26.7)**	**42.0**
		CardiacPhase [20]	5.10 (4.89)	4.70 (5.10)	189.6 (178.5)	174.7 (184.1)	182.1
Overall (view-macro avg.)		**Ours**	**2.29 (0.79)**	**2.05 (0.48)**	**42.9 (14.3)**	**39.5 (8.6)**	**41.2**
		CardiacPhase [20]	4.18 (4.54)	3.25 (3.94)	166.3 (181.4)	125.9 (146.8)	146.1

(2) CAMEO, to assess generalization beyond a single standardized view, we additionally evaluate on CAMEO, a multi-view echocardiography dataset containing 9 views per patient [11]. We split CAMEO into 8:1:1 for training, validation, and testing at the patient level.

3.1 Implementation Details

Frames are resized to 128×128 and normalized to $[0, 1]$. For training, we sample clips of length T=64 and temporally downsample by a factor of 2. The CNN is trained self-supervised with AdamW (learning rate 10^{-3}, weight decay 0.01, batch size 16) using cosine annealing and early stopping on validation loss. Loss weights are λ_{spec}=1.0, λ_{phase}=1.0, and λ_{cov}=10.0; the coverage log-barrier uses τ=0.1, β=0.1, and ϵ=10^{-8} (also used for numerical stability in divisions/logs). All experiments were conducted on a single RTX 3090 GPU. At inference, we apply the mask to full-length videos at native rate and generate ED/ES candidates via motion-energy, harmonic reconstruction, and additive fusion (Sec. 2.1–2.3). Being frequency-anchored, the method is window-based, not frame-causal: it

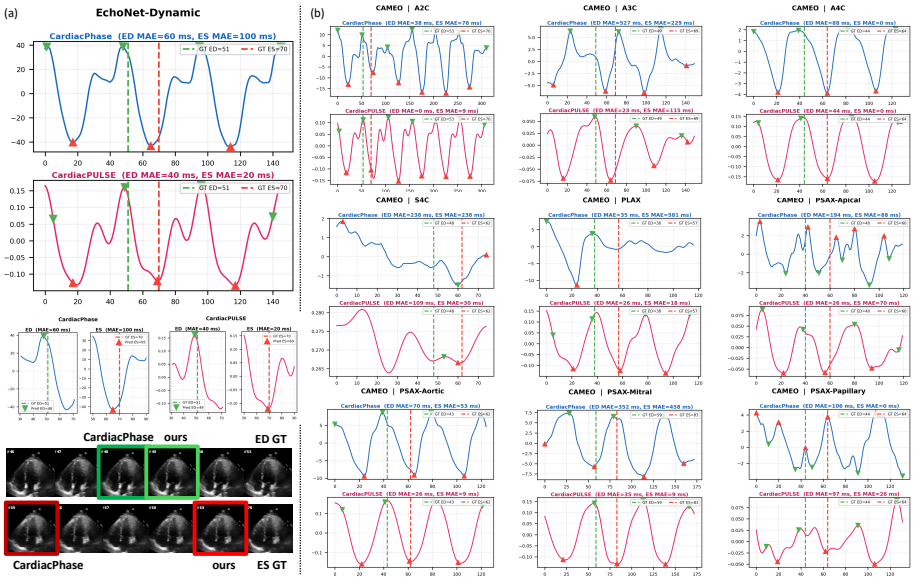

Fig. 2: Qualitative visualization of EchoNet-Dynamic [15] and CAMEO [11]. **(a)** EchoNet-Dynamic example comparing CardiacPhase and CardiacPULSE; dashed lines indicate GT ED/ES and markers denote predictions, with selected ED/ES frames shown below. **(b)** CAMEO multi-view examples showing predicted waveforms and ED/ES markers for CardiacPhase and CardiacPULSE.

needs $\geq$1 stable cycle (preferably two), with rolling-window aggregation for very short clips.

3.2 Main Results

Quantitative Results. We compare against supervised baselines on EchoNet-Dynamic [15] (R3D-50 [12], MAEF [23], UVT [17]) and annotation-free baselines (DDSB [1], CardiacPhase [20], RepNet-Row/RepNet-Lag [5]). In Table 1, CardiacPULSE achieves **34.4** ms MAE for ED and **33.6** ms MAE for ES (**34.0** ms Avg. MAE), improving over CardiacPhase (49.05 ms Avg. MAE) while using only **103K** trainable parameters. Phase-specific harmonic anchors provide view-consistent extrema that help disambiguate ED versus ES, reducing the swapping ambiguity that can arise when phase is recovered from unlabeled turning points or latent geometry. This is particularly important in multi-view CAMEO [11], where view-dependent appearance and contrast may invert the polarity of the intensity waveform. For CardiacPhase, we therefore select between the two peak–trough assignments (peaks$\leftrightarrow$ED vs. peaks$\leftrightarrow$ES) on the validation split for each view and keep the choice fixed at test time (Table 2). Our fusion of a coverage-first motion-energy candidate pool with harmonic anchors achieves **41.2** ms view-macro Avg. MAE, compared to 146.1 ms for CardiacPhase. As neither dataset

Table 3: Ablation[1] on spatial masks.

M_{cnn}	M_{fft}	ED	ES	Avg.
✗	✗	50.4	38.0	44.2
✗	✓	51.2	33.3	42.2
✓	✗	35.5	34.1	34.8
✓	✓	**34.4**	**33.6**	**34.0**

Table 4: Ablation[1] on self-supervised losses.

$\mathcal{L}_{\text{spec}}$	$\mathcal{L}_{\text{phase}}$	$\mathcal{L}_{\text{cov}}$	ED	ES	Avg.
✗	✓	✓	51.1	33.3	42.2
✓	✗	✓	36.3	33.9	35.1
✓	✓	✗	38.5	33.7	36.1
✓	✓	✓	**34.4**	**33.6**	**34.0**

Table 5: Ablation[1] on candidates generation.

$\mathcal{C}_{\text{ME}}$	f_0	$2f_0$	$3f_0$	$\mathcal{C}_{\text{ES}}^{\text{fft}}$	ED	ES	Avg.
✓	✗	✗	✗	✗	58.8	38.4	48.6
✓	✓	✗	✗	✗	46.3	38.4	42.4
✓	✓	✓	✗	✗	35.3	38.4	36.9
✓	✓	✓	✓	✗	34.4	38.4	36.4
✓	✓	✓	✓	✓	**34.4**	**33.6**	**34.0**
✗	✓	✓	✓	✗	103.2	38.4	70.8
Oracle ED					3.3	38.4	20.9

provides a heart-rate annotation, the cardiac cycle is derived from our self-estimated fundamental frequency f_0; a lightweight change-point selection then reduces the deduplicated pool to ≈1.7 ED / 1.5 ES candidates per cycle. Even at this sparsity, random same-count candidates exceed 90 ms, confirming accuracy reflects placement.

Qualitative Results. Fig. 2 suggests that CardiacPhase is vulnerable to view variability because ED/ES are inferred indirectly from geometric extrema in a learned latent trajectory, so phase assignment can drift when appearance, contrast, or non-cardiac motion changes, producing larger ED offsets and occasional large-error outliers in CAMEO. In contrast, CardiacPULSE imposes an explicit periodicity prior by extracting signals from cardiac-frequency regions and uses phase-specific harmonic anchors fused with a high-recall motion-energy pool, which stabilizes ED/ES correspondence across views, consistent with Table 2.

3.3 Ablation on Candidate Generation and Fourier-guided Masks

Spatial mask (Table 3). The mask is not used for segmentation; it is learned to denoise the extracted signals by suppressing background pixels that violate cardiac periodicity. M_{fft} alone is noisy, while M_{cnn} learned under spectral/phase objectives provides most of the gain; multiplying with M_{fft} further suppresses residual oscillators. The benefit is mainly on ED, which relies on a clean intensity waveform for harmonic anchoring.

Self-supervised loss (Table 4). All objectives are complementary: $\mathcal{L}_{\text{spec}}$ enforces a single dominant cardiac frequency in the extracted signal, $\mathcal{L}_{\text{phase}}$ promotes coherent oscillations within the selected region, and $\mathcal{L}_{\text{cov}}$ prevents degenerate masks by maintaining sufficient support for stable signal extraction.

Candidate generation (Table 5). Motion-energy turning points provide a high-recall but phase-agnostic pool, leading to ED ambiguity when used alone. Harmonic anchors add phase semantics by reconstructing a cycle-consistent waveform from $s_{\text{int}}(t)$: f_0 and especially $2f_0$ sharpen ED detection, while ES changes little until trough anchors are added since ES is already well captured by motion-energy minima. Fourier-only candidates lack coverage, confirming the need for motion energy.

4 Conclusion

In this work, we presented **CardiacPULSE**, a lightweight, self-supervised framework for ECG-free cardiac phase detection in echocardiography. By integrating cardiac periodicity as an explicit frequency prior via Fourier-guided masking and harmonic anchoring, **CardiacPULSE** localizes ED and ES frames without ECG, phase labels, or segmentation supervision. Extensive experiments on EchoNet-Dynamic and CAMEO demonstrate high accuracy and multi-view robustness using only 103K parameters. CardiacPULSE is an annotation-free front-end providing sparse ED/ES candidates for downstream LV-volume/EF pipelines, complementary to segmentation rather than replacing it. Its periodicity prior is also a limitation: severe arrhythmia, large view drift, probe motion, or heavy artifacts can blur the dominant frequency and remain failure modes. Future work will improve robustness to these distortions and extend to downstream quantification.

Acknowledgements. This work was supported by a research grant from the Joint Research Scheme (JRS) under the National Natural Science Foundation of China (NSFC) and the Research Grants Council (RGC) of Hong Kong (Project No. N_HKUST654/24); a research grant from the Innovation and Technology Commission (ITC) (Project No. GHP/124/22); and research grants from the RGC (Project Nos. 16500825, R6005-24, and C6088-25Y).

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

References

1. Bu, Z., Liu, Y., Huo, J., Peng, J., Wang, K., Zhou, G., Sparks, R., Dasgupta, P., Granados, A., Ourselin, S.: Ddsb: An unsupervised and training-free method for phase detection in echocardiography. In: International Workshop on Machine Learning in Medical Imaging. pp. 42–51. Springer (2024)
2. Chamsi-Pasha, M.A., Sengupta, P.P., Zoghbi, W.A.: Handheld echocardiography: current state and future perspectives. Circulation **136**(22), 2178–2188 (2017)
3. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. Nature Medicine **30**(5), 1481–1488 (2024)
4. Du, Y., Guo, J., Li, X.: Cardiacclip: Video-based clip adaptation for lvef prediction in a few-shot manner. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 46–56. Springer (2025)
5. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Counting out time: Class agnostic video repetition counting in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10387–10396 (2020)
6. Fadnavis, S., Parmar, C., Emaminejad, N., Ulloa Cerna, A., Malik, A., Selej, M., Mansi, T., Dunnmon, P., Yardibi, T., Standish, K., et al.: Echofm: a view-independent echocardiogram model for the detection of pulmonary hypertension. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 253–263. Springer (2024)

7. Farhad, M., Masud, M.M., Beg, A., Ahmad, A., Ahmed, L.A., Memon, S.: Cardiac phase detection in echocardiography using convolutional neural networks. Scientific reports **13**(1), 8908 (2023)
8. Ferraz, S., Coimbra, M., Pedrosa, J.: Assisted probe guidance in cardiac ultrasound: A review. Frontiers in Cardiovascular Medicine **10**, 1056055 (2023)
9. Holste, G., Oikonomou, E.K., Tokodi, M., Kovács, A., Wang, Z., Khera, R.: Complete ai-enabled echocardiography interpretation with multitask deep learning. JAMA **334**(4), 306–318 (2025)
10. Kong, B., Zhan, Y., Shin, M., Denny, T., Zhang, S.: Recognizing end-diastole and end-systole frames via deep temporal regression network. In: International conference on medical image computing and computer-assisted intervention. pp. 264–272. Springer (2016)
11. Lee, S.Y., Lei, K.Z., Chen, J.Y., Huang, C.R.: Descriptor: Cardiac multi-view echo dataset (cameo). IEEE Data Descriptions **3**, 47–53 (2025)
12. Li, Y., Li, H., Wu, F., Luo, J.: Semi-supervised learning improves the performance of cardiac event detection in echocardiography. Ultrasonics **134**, 107058 (2023)
13. Michailovich, O.V., Tannenbaum, A.: Despeckling of medical ultrasound images. ieee transactions on ultrasonics, ferroelectrics, and frequency control **53**(1), 64–78 (2006)
14. Mitchell, C., Rahko, P.S., Blauwet, L.A., Canaday, B., Finstuen, J.A., Foster, M.C., Horton, K., Ogunyankin, K.O., Palma, R.A., Velazquez, E.J.: Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the american society of echocardiography. Journal of the American Society of Echocardiography **32**(1), 1–64 (2019)
15. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al.: Video-based ai for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (2020)
16. Qin, Y., Wang, L., Zhao, C., Lee, A.P., Li, X.: Incentivizing cardiologist-like reasoning in mllms for interpretable echocardiographic diagnosis. arXiv preprint arXiv:2601.08440 (2026)
17. Reynaud, H., Vlontzos, A., Hou, B., Beqiri, A., Leeson, P., Kainz, B.: Ultrasound video transformers for cardiac ejection fraction estimation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 495–505. Springer (2021)
18. Vukadinovic, M., Tang, X., Yuan, N., Cheng, P., Li, D., Cheng, S., He, B., Ouyang, D.: Echoprime: a multi-video view-informed vision-language model for comprehensive echocardiography interpretation. arXiv preprint arXiv:2410.09704 (2024)
19. Yang, J., Lin, Y., Pu, B., Guo, J., Xu, X., Li, X.: Cardiacnet: Learning to reconstruct abnormalities for cardiac disease assessment from echocardiogram videos. In: European Conference on Computer Vision. pp. 293–311. Springer (2024)
20. Yang, Y., Yang, Q., Cui, K., Peng, C., D'Alberti, E., Hernandez-Cruz, N., Patey, O., Papageorghiou, A.T., Noble, J.A.: Latent motion profiling for annotation-free cardiac phase detection in adult and fetal echocardiography videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 316–325. Springer (2025)
21. Yue, Y., Wang, Y., Jiang, H., Liu, P., Song, S., Huang, G.: Echoworld: Learning motion-aware world models for echocardiography probe guidance. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25993–26003 (2025)

22. Zamzmi, G., Rajaraman, S., Hsu, L.Y., Sachdev, V., Antani, S.: Real-time echocardiography image analysis and quantification of cardiac indices. Medical image analysis **80**, 102438 (2022)
23. Zeng, Y., Tsui, P.H., Pang, K., Bin, G., Li, J., Lv, K., Wu, X., Wu, S., Zhou, Z.: Maef-net: Multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography. Ultrasonics **127**, 106855 (2023)
24. Zhang, J., Gajjala, S., Agrawal, P., Tison, G.H., Hallock, L.A., Beussink-Nelson, L., Lassen, M.H., Fan, E., Aras, M.A., Jordan, C., et al.: Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. Circulation **138**(16), 1623–1635 (2018)
25. Zolgharni, M., Negoita, M., Dhutia, N.M., Mielewczik, M., Manoharan, K., Sohaib, S.A., Finegold, J.A., Sacchi, S., Cole, G.D., Francis, D.P.: Automatic detection of end-diastolic and end-systolic frames in 2d echocardiography. Echocardiography **34**(7), 956–967 (2017)