Motion-Conditioned Multi-View Fusion for Myocardial Infarction Localization from Echocardiography

Guang Yang[1⋆], Wentian Xu[1], Siyu Wang[1], Betty Raman[2], Lei Li[3], and Vicente Grau[1]

[1] Department of Engineering Science, University of Oxford, Oxford, United Kingdom
guang.yang@eng.ox.ac.uk
[2] Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom
[3] Department of Biomedical Engineering, National University of Singapore, Singapore

Abstract. Myocardial infarction (MI) remains a leading cause of mortality worldwide. Echocardiography (Echo) is a widely available modality for MI assessment, where regional wall motion abnormality is a key indicator. Prior learning based methods for myocardial motion analysis often use handcrafted descriptors or densely supervised estimation, but the need for extensive annotation limits applicability. Foundation models have recently improved vision-based Echo analysis; however, most methods operate on single views and segment-level localization remains unreliable under view-dependent ambiguity, especially in apical views. To address this, we propose MCF-Net, a novel motion-guided multi-view fusion framework that fuses myocardial motion cues with foundation model representations to localize infarction. Visual features are extracted using EchoPrime, a pretrained Echo foundation model shared across dual views. Cardiac motion is modeled with extremely sparse supervision: a single annotated template frame is transferred across videos to initialize point tracking, avoiding dense labels. Motion-derived segment-aware soft masks provide coarse spatial priors that selectively enhance features for challenging myocardial segments. A motion-conditioned fusion mechanism then integrates motion and vision across views, refining predictions without overriding strong appearance cues. On segment-level MI localization, MCF-Net achieves 72.4% F1 and 84.9% accuracy, outperforming state-of-the-art motion-only, vision-only, and fusion baselines. Code is available at `https://github.com/GYang1619/MCF-Net-MICCAI2026`.

Keywords: Myocardial infarction · Echocardiography · Multi-view fusion · Motion priors

1 Introduction

Myocardial infarction (MI) refers to irreversible myocardial damage caused by coronary artery occlusion. The associated mechanical dysfunction often results

⋆ Corresponding author.

in abnormal myocardial motion [23]. Related contrast-free work in cine MRI has also used cardiac motion to reconstruct patient-specific infarct geometry [10], supporting the value of cardiac mechanics for infarct assessment. Echocardiography (Echo), with its high temporal resolution, is routinely used to assess MI via regional wall motion abnormality (RWMA). In clinical practice, MI assessment integrates multiple acquisition views; the assessment is most commonly made in apical four- and two-chamber (A4C/A2C) views, which provide different coverage of the left ventricle. Following the AHA 17-segment model [3], each apical view maps to six distinct myocardial segments. However, Echo often suffers from limited image quality and view-dependent ambiguity, making RWMA assessment highly dependent on clinician expertise [14] and hindering scalable, reproducible MI localization.

Most automated approaches model RWMA through explicit motion analysis. Optical flow based methods estimate pixel wise motion without anatomical labels but are sensitive to noise and imaging artifacts, often yielding physically implausible motion fields [11, 16]. Image registration methods estimate dense deformation between frames, yet their performance degrades without accurate anatomical supervision [1, 2, 21]. Biomechanical and deformation based models further rely on detailed anatomical annotations to derive physiological parameters, resulting in substantial labeling burden [6, 15, 17, 22]. In contrast, we model cardiac motion using an extremely sparse point tracking strategy that requires only a single annotated template frame for the entire dataset. We transfer this template to initialize landmarks in each video and propagate them over time using a tracking model [7], producing dense trajectories with minimal supervision.

Foundation models (FMs) have recently become a dominant paradigm in medical image analysis, driven by large scale pretraining and contrastive learning, leading to growing interest in MI detection from Echo [4, 8, 20]. Recent work has also explored anatomical significance and multi-task learning for explainable MI prediction [12]. EchoPrime [20], trained on over 12 million image-text pairs and evaluated across 23 cardiac benchmarks, provides a high capacity visual backbone for routine tasks. However, despite reliable global predictions, our experiments indicate that EchoPrime remains limited for segment level MI localization, particularly in structurally ambiguous regions in the A4C view. This motivates utilizing cardiac motion as a complementary cue to guide visual attention toward pathology relevant regions.

Motivated by the complementary strengths and inherent limitations of motion and vision based approaches, we propose **MCF-Net**, a Motion Conditioned Fusion Network for segment level MI localization. MCF-Net adopts a pretrained foundation model as the primary vision encoder and integrates motion guidance through two components: (i) motion-guided soft refinement, which constructs a trajectory-derived Gaussian mask along the tracked myocardial boundary to lead the encoder towards relevant regions, and (ii) motion conditioned multi-view fusion, which uses motion embeddings to gate visual features and applies cross-view attention to aggregate complementary evidence from both views. MCF-Net consistently outperforms all motion only methods (trajectory/motion features

only), vision only methods (visual features only), and previous fusion baselines, including strong FMs, achieving up to +8.1% F1 and +4.8% accuracy gains for segment MI localization. Our main contributions are:

- We introduce a motion conditioned multi-view localization framework that combines a pretrained echocardiography foundation model with sparse motion priors to enable segment level MI localization across apical views.
- We design a motion guided refinement module (soft mask) and a motion conditioned fusion strategy that selectively pushes visual representations toward pathology relevant regions, improving hard myocardial segments (such as basal regions) while preserving strong global appearance cues.
- We propose a sparse motion modeling strategy using a single annotated frame with template transfer and point tracking, enabling robust and scalable cardiac motion extraction across views without the need for dense labels.

2 Method

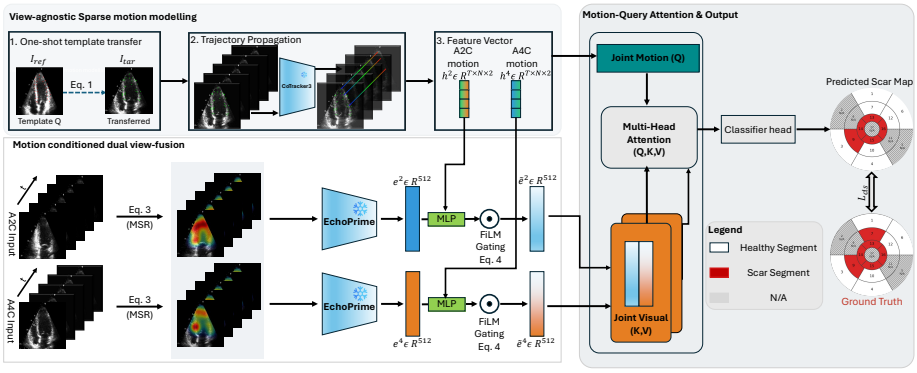

Fig. 1. Overview of MCF-Net. (1) **Sparse Motion Modeling** extracts motion embeddings (h^2, h^4) via CoTracker3. (2) **Dual View Fusion** modulates frozen EchoPrime visual features with motion via FiLM gating (Eq. 4). (3) **Motion Query Attention** fuses representations using Joint Motion as Q and Visual as K, V to predict 12 segment MI distribution, optimized via $\mathcal{L}_{cls}$.

Given a paired dual view Echo $\mathcal{X} = \left(\mathcal{V}^{(\mathrm{A2C})}, \mathcal{V}^{(\mathrm{A4C})}\right)$, where each view is represented as a temporal sequence $\mathcal{V}^{(v)} = \{I_t^{(v)} \in \mathbb{R}^{H\times W}\}_{t=1}^{T}$ with $v \in \{\mathrm{A2C}, \mathrm{A4C}\}$, our goal is segment-level MI localization, learning $f_\theta : \mathcal{X} \to \mathbf{z} \in \mathbb{R}^{12}$ with $\mathbf{z}$ the segment-wise logits. As shown in Fig. 1, **MCF-Net** integrates sparse motion priors with foundation appearance features via: (i) sparse motion modeling, (ii) motion-guided refinement, and (iii) motion-conditioned cross-view fusion.

2.1 Cardiac Motion Modeling from Extremely Sparse Annotation

To avoid the extensive cost of dense frame by frame human annotation, we extract cardiac motion under extremely sparse supervision. Specifically, our approach requires only one manually annotated frame for the entire dataset: we annotate the canonical landmarks on the first frame of a single reference A4C video and initialize landmarks for all videos via localized template matching followed by point tracking. When switching to A2C view, we directly transfer the A4C template across views using the same matching and tracking pipeline. Importantly, A2C and A4C views are initialized and tracked independently. No explicit point-to-point anatomical correspondence between views is assumed. Formally, we define a canonical landmark set $\mathcal{Q} = \{q_i\}_{i=1}^{N}$ with $N = 32$ on the first frame of a reference A4C video $\mathcal{I}_{\text{ref}}$, where the points are sampled along the myocardial wall with approximately uniform allocation across the six view-specific segments. For each landmark q_i, we extract a feature patch $\mathcal{P}_i \in \mathbb{R}^{11\times 11}$ centered at q_i. To transfer these landmarks to the first frame of a target video $\mathcal{I}_{\text{tar}}$, we perform localized patch matching within a search window $\Omega \in \mathbb{R}^{41\times 41}$ centered at the same spatial coordinate q_i. Correspondence between anatomical features is established by maximizing the Normalized Cross Correlation (NCC):

$$\text{NCC}(\mathbf{u}) = \frac{\sum_{\mathbf{x}\in\mathcal{P}_i}(\mathcal{I}_{\text{ref}}(\mathbf{x}) - \bar{\mathcal{I}}_{\text{ref}})(\mathcal{I}_{\text{tar}}(\mathbf{x}+\mathbf{u}) - \bar{\mathcal{I}}_{\text{tar}})}{\sqrt{\sum_{\mathbf{x}}(\mathcal{I}_{\text{ref}}(\mathbf{x}) - \bar{\mathcal{I}}_{\text{ref}})^2 \sum_{\mathbf{x}}(\mathcal{I}_{\text{tar}}(\mathbf{x}+\mathbf{u}) - \bar{\mathcal{I}}_{\text{tar}})^2}}, \quad \mathbf{u} = (u_x, u_y) \in \Omega. \tag{1}$$

To ensure our method can robustly locate $P = \{p_i(1)\}_{i=1}^{N}$ against acoustic artifacts and structural variations, we employ a high confidence fallback mechanism. The initialized point p_i in $\mathcal{I}_{\text{tar}}$ is defined as follows:

$$p_i = q_i + \arg\max_{\mathbf{u}\in\Omega} \text{NCC}(\mathbf{u}). \tag{2}$$

If $\max_{\mathbf{u}\in\Omega} \text{NCC}(\mathbf{u}) < \tau$, we set $p_i = q_i$, where $\tau = 0.5$. This fallback ensures that if no reliable structural match is found (e.g., due to poor echo quality), the coordinate defaults to the atlas prior q_i, maintaining a structurally plausible heart shape. The initialized landmarks $\{p_i(1)\}_{i=1}^{N}$ are propagated across T frames using CoTracker3 [7], producing dense $\mathbf{P} = \{p_i(t)\}_{t=1,i=1}^{T,N} \in \mathbb{R}^{T\times N\times 2}$.

2.2 Foundation Encoding with Motion Guided Soft Refinement

While FMs such as EchoPrime provide strong spatial representations, they often lack temporal specificity for infarction detection, where subtle wall thinning and hypokinesia along the myocardial boundary are key biomarkers. To bridge this gap, we introduce **Motion Guided Soft Refinement (MSR)**, which injects tracked trajectories into the appearance stream. Since $\mathcal{Q}$ is sampled along the myocardial wall with roughly equal points per segment, the resulting trajectories $\{p_i(t)\}_{i=1}^{N}$ induce a segment-aware spatial prior. Therefore we construct a spatial-temporal soft attention mask $M^{(v)}(t, x, y)$ using a Gaussian mixture centered at

the landmark locations: $M^{(v)}(t,x,y) = \sum_{i=1}^{N} \exp\Big(-\frac{\|(x,y)-p_i(t)\|_2^2}{2\sigma^2}\Big)$. This mask highlights myocardial regions and is used to softly modulate the input video $\mathcal{V}^{(v)}$ as follows:

$$\tilde{\mathcal{V}}^{(v)} = \mathcal{V}^{(v)} \odot \big(1 + \lambda M^{(v)}\big), \tag{3}$$

where λ controls attention strength and $\odot$ denotes element wise multiplication. This operation enhances intensity and edge responses around the myocardial wall while largely preserving the original appearance distribution and surrounding context (e.g., the left ventricular cavity and surroundings). The refined video is then fed into the frozen foundation encoder f_ϕ to extract view level embeddings $e^{(v)} = f_\phi\Big(\tilde{\mathcal{V}}^{(v)}\Big) \in \mathbb{R}^{512}$. Overall, this refinement retains the pretrained spatial priors of f_ϕ while injecting localized motion guidance, improving infarction sensitive feature extraction without altering the encoder parameters.

2.3 Motion Conditioned Cross View Fusion

MI localization benefits from complementary evidence across both views, since different AHA segments are best visualized under different orientations. To exploit this complementarity, we propose a motion conditioned fusion module that explicitly couples visual representations with cardiac motion. For each view v, the frozen foundation encoder outputs a visual embedding $e^{(v)} \in \mathbb{R}^{512}$, while the tracked motion trajectory $\mathbf{P}^{(v)} \in \mathbb{R}^{T\times N\times 2}$ is summarized by an MLP into a compact motion descriptor $h^{(v)} \in \mathbb{R}^{128}$. We then modulate appearance features via FiLM style conditioning [13]:

$$\tilde{e}^{(v)} = \gamma\Big(h^{(v)}\Big) \odot e^{(v)} + \beta\Big(h^{(v)}\Big) \in \mathbb{R}^{512}, \tag{4}$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ are learned linear mappings. This conditioning enables motion abnormalities (e.g., reduced contraction patterns) to re-weight the visual embedding in a view-specific manner, promoting infarction-relevant cues while suppressing non informative variations. To fuse across two views, we concatenate the modulated embeddings $V = [\tilde{e}^{(\mathrm{A2C})}, \tilde{e}^{(\mathrm{A4C})}]$ and form a joint motion query $q = [\, h^{(\mathrm{A2C})}; h^{(\mathrm{A4C})}\,]$. Cross view fusion is performed using multi-head attention, $z = \mathrm{MHA}(q,V,V) \in \mathbb{R}^{512}$, so that pooling across views is explicitly guided by patient-specific motion. Finally, the fused representation is mapped to segment-wise infarction probabilities $\hat{y} = \sigma(W_o z) \in (0,1)^{12}$. The training objective is optimized by segment-wise classification loss $L_{\mathrm{cls}} = L_{\mathrm{BCE}}(y,\hat{y})$. Overall, the proposed fusion couples motion-informed inter-view modulation with attention-based cross-view aggregation, enabling view-complementary and pathologically-informed representations that improve segment level MI localization.

3 Experiments

3.1 Materials

Dataset This study uses the HMC-QU dataset [5], the only publicly available Echo dataset for MI detection to date. HMC-QU comprises 162 A4C and 160

A2C Echo videos, with paired A2C and A4C views available for 160 patients. Following the standardized AHA 17-segment model [3], MI presence is annotated for the six segments visible in each view. We randomly partitioned the data into 112/16/32 patients for training, validation, and testing. All experiments were repeated three times, with performance reported as segment-level averages.

Implementation All experiments were conducted on a workstation equipped with an Intel i7-13700 CPU and an NVIDIA RTX A4000 GPU. Echo videos were preprocessed by resizing frames to 224 × 224 pixels and uniformly downsampled to 16 frames to meet the input requirements of EchoPrime. The proposed framework was implemented in PyTorch. Model training employed the AdamW optimizer with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}. All models were trained for 50 epochs with a batch size of 16.

3.2 Results

We evaluate segment-level MI localization using threshold-free (AUROC, PR-AUC) and threshold-dependent (F1, ACC) metrics, averaged over three runs. MCF-Net is compared against seven baselines spanning motion-only [5,9], vision-only [18–20], and fusion-based methods, including naïve EchoPrime fusion and EchoAna [22]. All models operate on full echo videos, enabling systematic comparison of motion priors, foundation representations, and fusion strategies.

Table 1. Quantitative comparison of segment level MI localization. Experiments are run three times with different seeds. Results are reported as mean ± standard deviation. Best results are **bold**. PR-AUC denotes the area under the precision–recall curve.

Methods	Motion	Vision	AUROC	PR-AUC	F1-Score	Accuracy
APs [9]	✓	✗	–	–	57.7 ± 5.8	79.0 ± 2.6
APsML [5]	✓	✗	83.0 ± 2.1	70.4 ± 7.2	59.9 ± 6.0	79.8 ± 1.7
InceptionV3 [18]	✗	✓	61.1 ± 8.0	47.2 ± 12.0	27.0 ± 2.8	61.5 ± 1.1
R(2+1)D [19]	✗	✓	77.3 ± 1.7	58.2 ± 3.9	45.6 ± 9.8	73.3 ± 5.4
EchoPrime [20]	✗	✓	84.9 ± 2.6	72.7 ± 6.2	64.3 ± 3.8	80.1 ± 2.1
EchoPrimeNF [20]	✓	✓	84.4 ± 1.8	70.5 ± 6.2	59.3 ± 5.6	78.8 ± 1.7
EchoAna [22]	✓	✓	84.6 ± 1.9	72.1 ± 6.5	61.7 ± 4.6	80.9 ± 2.8
Ours	✓	✓	**87.6 ± 2.4**	**74.7 ± 5.6**	**72.4 ± 3.8**	**84.9 ± 2.3**

Comparison Studies Table 1 shows that MCF-Net achieves the best performance across all metrics for segment level MI localization (AUROC 87.6, PR-AUC 74.7, F1 72.4, ACC 84.9). Improvements are consistent in both threshold-free (AUROC/PR-AUC) and threshold-dependent (F1/ACC) metrics, indicating gains in discrimination and region level decision quality. **Motion only methods**

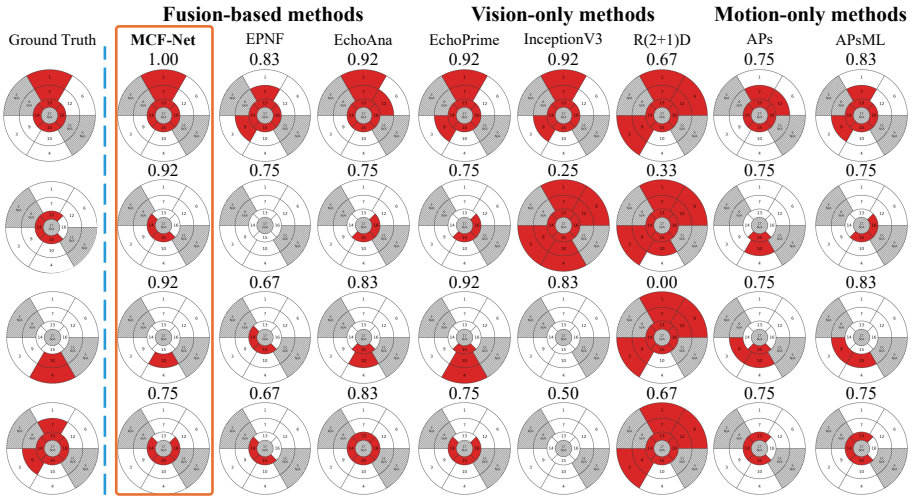

Fig. 2. Qualitative comparison on AHA bullseye plots. Each "segment" denotes a standardized *anatomical region* of the left ventricle (apical/mid/basal; anterior/inferior, etc.), i.e., region-level labeling rather than pixel-wise segmentation. **Red, white, gray** denote infarct, healthy, and N/A regions, respectively. Rows show representative test cases selected by accuracy quartiles (0.75: lower, 0.92: median, 1.00: upper), comparing Ground Truth (GT) with **MCF-Net (Ours)** and baselines grouped by modality (Fusion, Vision, Motion). Numbers indicate accuracy.

(APs, APsML) achieve reasonable overall accuracy (~ 80) but lack spatial precision, resulting in limited segment level localization (APsML: PR-AUC 70.4 ± 7.2, F1 59.9 ± 6.0). **Vision only models** show that standard backbones struggle (InceptionV3 F1 27.0 ± 2.8, R(2+1)D F1 45.6 ± 9.8), while the pretrained foundation model EchoPrime substantially improves performance (AUROC 84.9 ± 2.6, F1 64.3 ± 3.8), highlighting the benefit of large scale pretraining. **Fusion methods prove critical.** Naïve fusion (EchoPrimeNF) degrades performance relative to EchoPrime (F1 59.3 ± 5.6 vs. 64.3 ± 3.8), suggesting misaligned motion-vision interaction. In contrast, our motion-conditioned fusion consistently improves all metrics, demonstrating the benefit of conditioning visual representations before cross view aggregation. In addition, paired patient-level bootstrap analysis confirms the F1 improvement of MCF-Net over EchoPrime (Delta=0.08, 95% CI=[0.01,0.15], $p<0.05$), indicating that the observed performance gain is statistically significant. Qualitative results in Fig. 2 further support these findings. Across representative cases spanning lower to upper performance quartiles, MCF-Net produces compact and anatomically coherent infarct regions. Vision only methods often overestimate lesions (diffuse false positives), while motion only approaches generate fragmented patterns. Alternative fusion strategies may shift activation to adjacent segments. In contrast, MCF-Net more accurately recovers infarct extent and topology, consistent with its quantitative gains.

Table 2. Ablation study of Motion Conditioned Fusion (MCF) and Motion Guided Soft Refinement (MSR). Results are reported as mean ± standard deviation.

Variant	MCF	MSR	AUROC	PR-AUC	F1-Score	Accuracy
Baseline	✗	✗	84.9 ± 2.6	72.7 ± 6.2	64.3 ± 3.8	80.1 ± 2.1
Baseline + MSR	✗	✓	86.9 ± 2.8	71.9 ± 6.7	66.8 ± 3.6	82.4 ± 3.6
Baseline + MCF	✓	✗	87.2 ± 2.5	**74.7 ± 5.8**	70.3 ± 4.8	83.8 ± 3.3
Full Model	✓	✓	**87.6 ± 2.4**	74.6 ± 5.6	**72.4 ± 3.8**	**84.9 ± 2.3**

Ablation Study As shown in Table 2, the baseline model (EchoPrime encoder + linear head) achieves strong overall performance (AUROC 84.9 ± 2.6, F1 64.3 ± 3.8), reflecting the benefit of large scale pretraining for global representations, yet its gains are less pronounced when segment level MI localization is required. Adding MSR yields a modest but consistent improvement (AUROC 86.9 ± 2.8, F1 66.8 ± 3.6), suggesting that the soft mask mainly sharpens the decision boundary by steering features towards the myocardial region. Incorporating MCF provides a larger boost (F1 70.3 ± 4.8, ACC 83.8 ± 3.3, PR-AUC 74.7 ± 5.8), indicating increased model capacity to capture motion appearance interactions that are critical for segment level reasoning. Combining both modules achieves the best results (AUROC 87.6 ± 2.4, F1 72.4 ± 3.8, ACC 84.9 ± 2.3), confirming their complementarity: MCF strengthens cross domain fusion, while MSR refines spatial emphasis to improve separability and localization robustness.

The Effect of MSR Strength λ We analyze the impact of λ in Equation 3, which scales the motion derived soft mask. As shown in Fig. 3, performance follows a clear trend, with a moderate refinement strength ($\lambda = 0.3$) yielding the best overall performance (AUROC 87.6, F1 72.4, Acc 84.9). A stable high performance region is observed for $\lambda \in [0.2, 0.4]$, indicating robustness to mild hyperparameter variation. However, performance consistently degrades when $\lambda \geq 0.5$, with F1 decreasing to 68.3 at $\lambda = 1.0$, falling below the baseline. This suggests that excessive masking can distort the original Echo, lead to artefactual enhancements and ultimately degrade localization reliability. Notably, while Accuracy and F1 improve significantly at moderate λ, PR-AUC remains relatively stable, implying that soft refinement primarily improves decision boundaries rather than ranking quality. These findings support $\lambda = 0.3$ as an optimal trade-off between motion enhancement and visual fidelity.

4 Conclusion

In this work, we present MCF-Net, a Motion Conditioned Fusion Network for segment level myocardial infarction localization from dual view echocardiography. Our framework utilizes a pretrained foundation model to extract rich visual representations and introduces Motion Guided Soft Refinement to focus feature extraction on infarction-relevant regions. Motion Conditioned Fusion modulates

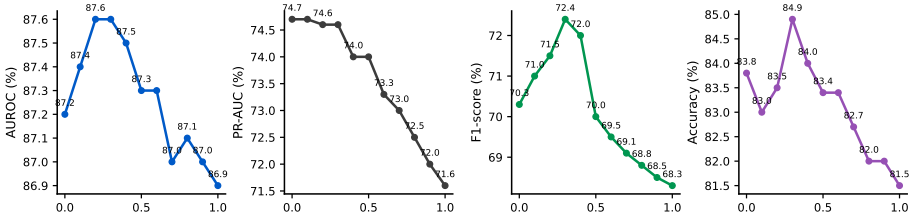

Fig. 3. Sensitivity analysis of the refinement strength λ on the test set (higher λ strengthens masking). From left to right: AUROC, PR-AUC, F1-score, and accuracy.

appearance features using view-specific motion and enables cross-view interaction between motion dynamics and visual semantics, improving fine grained localization. Experimental results validate that sparse physiological motion priors can further enhance strong foundation models such as EchoPrime, providing a practical approach for MI localization from routine echocardiography. Future work will investigate larger multi-center validation and systematic failure-case analysis.

Acknowledgments. The authors acknowledge the use of computing cluster resources provided by the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford.

Disclosure of Interests. The authors declare no competing interests.

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
2. Bi, N., Zakeri, A., Xia, Y., Cheng, N., Taylor, Z.A., Frangi, A.F., Gooya, A.: Segmorph: concurrent motion estimation and segmentation for cardiac mri sequences. IEEE transactions on medical imaging (2024)
3. Cerqueira, M., Weissman, N., Dilsizian, V., Jacobs, A.: Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association. Circulation **105**(4), 539–542 (2002)
4. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. Nature Medicine **30**(5), 1481–1488 (2024)
5. Degerli, A., Kiranyaz, S., Hamid, T., Mazhar, R., Gabbouj, M.: Early myocardial infarction detection over multi-view echocardiography. Biomedical Signal Processing and Control **87**, 105448 (2024)
6. Gomez, C., Letizia, A., Tufano, V., Molinari, F., Salvi, M.: A cascade approach for the early detection and localization of myocardial infarction in 2d-echocardiography. Medical Engineering & Physics **143**(1), 104400 (2025)

7. Karaev, N., Makarov, Y., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6013–6022 (2025)
8. Kim, S., Jin, P., Song, S., Chen, C., Li, Y., Ren, H., Li, X., Liu, T., Li, Q.: Echofm: Foundation model for generalizable echocardiogram analysis. IEEE transactions on medical imaging (2025)
9. Kiranyaz, S., Degerli, A., Hamid, T., Mazhar, R., Ahmed, R.E.F., Abouhasera, R., Zabihi, M., Malik, J., Hamila, R., Gabbouj, M.: Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection. IEEE Access **8**, 210301–210317 (2020)
10. Lyu, Y., Yang, F., Liu, X., Jiang, Z., Dillon, J., Zhao, D., Nash, M., Mauger, C., Young, A., Sia, C.H., et al.: Personalized 3d myocardial infarct geometry reconstruction from cine mri with explicit cardiac motion modeling. In: International Workshop on Digital Twin for Healthcare. pp. 1–11. Springer (2025)
11. Ortiz-Gonzalez, A., Kobler, E., Simon, S., Bischoff, L., Nowak, S., Isaak, A., Block, W., Sprinkart, A.M., Attenberger, U., Luetkens, J.A., et al.: Optical flow-guided cine mri segmentation with learned corrections. IEEE Transactions on Medical Imaging **43**(3), 940–953 (2023)
12. Peng, J., Beetz, M., Banerjee, A., Chen, M., Grau, V.: An anatomical significance-aware architecture for explainable myocardial infarction prediction via multi-task learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 13–23. Springer (2025)
13. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
14. Porter, T.R., Mulvagh, S.L., Abdelmoneim, S.S., Becher, H., Belcik, J.T., Bierig, M., Choy, J., Gaibazzi, N., Gillam, L.D., Janardhanan, R., et al.: Clinical applications of ultrasonic enhancing agents in echocardiography: 2018 american society of echocardiography guidelines update. Journal of the American Society of Echocardiography **31**(3), 241–274 (2018)
15. Qin, C., Wang, S., Chen, C., Qiu, H., Bai, W., Rueckert, D.: Biomechanics-informed neural networks for myocardial motion tracking in mri. In: International conference on medical image computing and computer-assisted intervention. pp. 296–306. Springer (2020)
16. Suhling, M., Arigovindan, M., Jansen, C., Hunziker, P., Unser, M.: Myocardial motion analysis from b-mode echocardiograms. IEEE Transactions on image processing **14**(4), 525–536 (2005)
17. Sundar, H., Davatzikos, C., Biros, G.: Biomechanically-constrained 4d estimation of myocardial motion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 257–265. Springer (2009)
18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
19. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
20. Vukadinovic, M., Chiu, I.M., Tang, X., Yuan, N., Chen, T.Y., Cheng, P., Li, D., Cheng, S., He, B., Ouyang, D.: Comprehensive echocardiogram evaluation with view primed vision language ai. Nature pp. 1–3 (2025)

21. Yang, G., Chen, J., Sheng, X., Yang, S., Zhuang, X., Raman, B., Li, L., Grau, V.: Contrast-free myocardial scar segmentation in cine mri using motion and texture fusion. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
22. Yang, Y., Rocher, M., Moceri, P., Sermesant, M., et al.: Echocardiography analysis with deep learning using priors: Multi-centric evaluation of generalisation. Machine Learning for Biomedical Imaging **2**(November 2024 issue), 2293–2325 (2024)
23. Zeidan, R.K., Farah, R.: Myocardial infarctions in developing countries. In: Handbook of Medical and Health Sciences in Developing Countries: Education, Practice, and Research, pp. 1–30. Springer (2024)