CycleSpectra: Cyclic Motion Spectra for Phase-Queryable 4D Cardiac Reconstruction

Xueming Fu[1,2]* **Xin Li**[1,2]* **Lixia Han**[6] **Ao Shen**[7]
Guangming Lu[8] **Song Luo**[9]† **Shaohua Kevin Zhou**[1,2,3,4,5]†

[1]School of Biomedical Engineering, Division of Life Sciences and Medicine, USTC [2]Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE) Lab, YRD-RIGHT, USTC Suzhou Institute for Advanced Research [3]Jiangsu Key Laboratory of Multimodal Digital Twin Technology, Suzhou, Jiangsu, China [4]Biomedical Basic Research Center (BBRC) of Jiangsu, Suzhou, Jiangsu, China [5]State Key Laboratory of Precision and Intelligent Chemistry, Suzhou, Jiangsu, China [6]Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China [7]Department of Radiology & Institute for Medical Imaging Technology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China [8]Department of Radiology, Jinling Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing, Jiangsu, China [9]Department of Radiology, Geriatric Hospital of Nanjing Medical University, Nanjing, Jiangsu, China

Abstract

Reconstructing patient-specific cyclic cardiac motion at arbitrary phases is central to dynamic cardiac assessment, underpinning measurements such as ejection fraction, stroke volume, and regional myocardial deformation. Existing methods—pair-wise registration, time-conditioned decoders, and coordinate-based implicit fields—share a common implicit assumption: the cardiac cycle is treated as a sequence of discrete phases, with temporal structure assembled through learned mappings supervised at the observed frames. Yet the cycle is a closed periodic process. We propose **CycleSpectra**, a spectral motion representation that takes the cyclic structure as a starting point of the representation rather than as a property to be learned. Given a pair of sparsely sampled phase observations, the model learns the spectral structure of the underlying cyclic cardiac motion in a low-dimensional latent space, and recovers the displacement at any phase through a fixed analytic synthesis layer—time never enters the network as a learnable signal, only as a coordinate the readout takes by construction. To support evaluation, we collect a thin-slice phase-resolved 4D cardiac CT cohort and benchmark under both canonical and non-canonical input protocols, reporting whole-cycle motion-fidelity metrics alongside standard image and anatomical metrics. Under arbitrary two-frame input, CycleSpectra attains the lowest mean left-ventricular ejection-fraction error and the highest mean ventricular volume–time curve correlation among recent baselines, while remaining competitive on image fidelity. Beyond two-frame inference, the proposed framework supports test-time multi-frame latent fusion for higher accuracy, validated on both the in-house cohort and a cross-modality 4D cardiac MR cohort. Code is available at `https://github.com/windrise/CycleSpectra`.

1 Introduction

Reconstructing the full cardiac cycle of motion from sparse observations is central to dynamic cardiac assessment, supporting clinical quantities such as left-ventricular ejection fraction (LVEF), stroke volume, and regional myocardial deformation [Morales et al., 2021, Qin et al., 2023]. The

*Equal contribution.
†Corresponding authors.

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

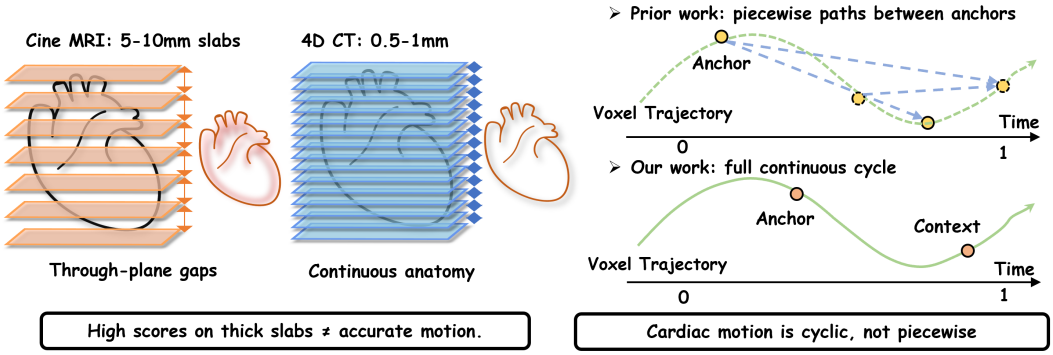

Figure 1: Left: thick cine slabs inflate image scores and hide true motion. Right: real voxel trajectories (green) are non-monotonic, but prior work only learns piecewise paths between anchors (blue).

available cardiac motion benchmarks shape this task in two subtle but consequential ways (Fig. 1). The dominant cohort, ACDC [Bernard et al., 2018], uses 5–10 mm short-axis slabs whose through-plane blurring lets image-quality metrics over-score reconstructions of inadequately recovered motion (left); and it annotates only end-diastole (ED) and end-systole (ES), encouraging methods to anchor learning at these two extrema and treat the cardiac cycle as a transition between anchors rather than as a process unfolding across phases (right).

Existing works can be summarized into two natural lines: pair-wise registration methods [Balakrishnan et al., 2019, Chen et al., 2022, 2024, Yang et al., 2024, Wang et al., 2025] that estimate a single deformation field between two discrete frames and time-aware methods [Krebs et al., 2019, Zhang et al., 2025a, Kim et al., 2024, Yuan et al., 2023, Fu et al., 2025] that bring temporal information into the model in various forms—scalar phase conditioning, learned latents, or coordinate-based fields. Despite the surface differences, these approaches share a common implicit assumption: the cardiac cycle is treated as a sequence of discrete phases, with the temporal structure assembled through learned mappings supervised at the observed frames. Yet the cardiac cycle is neither monotonic nor open-ended: it is a closed periodic process that the heart traces out beat after beat. *Why isn't cardiac motion's periodicity encoded in the representation itself?*

We take the cyclic structure as the starting point of the representation rather than as a property the model needs to learn. Concretely, given a pair of sparsely sampled phase observations, the model learns the spectral structure of their underlying cyclic cardiac motion in a low-dimensional latent space, and recovers the displacement at any phase through a fixed analytic synthesis layer rather than through repeated network forwards conditioned on time. Time never enters the network as a learnable signal—it is a coordinate the analytic readout takes by construction. This decouples the representation from any learned phase mapping, so periodicity and continuity hold by design. A direct consequence is that any known frame at any phase backward-supervises the same coefficient set, so the cyclic structure is built into the representation rather than memorized through the supervised frames. We design our framework around this representation, targeting full-cycle reconstruction from sparse phase observations—including the clinically realistic two-frame case where the inputs are not the canonical ED/ES anchors. Here "sparse" refers to the test-time conditioning (two frames); training uses dense per-phase query supervision (Sec. 4.2), and we ablate this in Sec. 4.4. To support evaluation in this regime, we collect a thin-slice phase-resolved 4D cardiac CT cohort and report whole-cycle motion-fidelity metrics (LVEF, LV/RV volume–time correlation) under both canonical ED/ES and non-canonical opposite-pair input protocols. Our contributions are summarized as follows:

1. We propose a spectral motion representation for cyclic cardiac dynamics, CycleSpectra, where the cyclic motion trajectory of each spatial site is parameterized by a small set of Fourier coefficients. Periodicity and continuity are built into the representation, and the deformation at any phase relative to the anchor is queryable from this representation.
2. We design a cyclic motion reconstruction framework that encodes the cyclic motion spectrum in a low-dimensional feature space from sparse phase observations, decodes per-site Fourier coefficient maps, and reads out the latent-scale flow analytically before refining it to the image-space

deformation. The framework also supports more input observations for more accurate motion spectrum prediction.

3. We collect a phase-resolved 4D cardiac CT cohort with thin-slice resolution and benchmark methods under both canonical ED/ES and non-canonical opposite-pair input protocols. CycleSpectra attains the best mean clinical motion-fidelity metrics (LVEF error and LV/RV volume–time correlation) under the opposite-pair protocol while remaining competitive on image fidelity. As the test set has seven subjects, we report subject-level bootstrap intervals and claim superiority only where the paired interval excludes zero (Sec. 4.3).

2 Related Work

Pair-wise registration. Pair-wise registration estimates the deformation field that maps one cardiac frame to another at a different time point. Building on classical diffeomorphic frameworks, learning-based methods such as VoxelMorph [Balakrishnan et al., 2019], TransMorph [Chen et al., 2022], Atlas-ISTN [Sinclair et al., 2022], TransMatch [Chen et al., 2024], and the GPTrack [Yang et al., 2024], ModeTV2 [Wang et al., 2025] have produced accurate frame-to-frame alignment with strong warped-image fidelity on annotated ED/ES pairs. These methods solve frame-to-frame alignment well but do not, by themselves, represent the cardiac cycle as a continuous process.

Time-aware methods. A second body of work brings temporal information into the model in three forms. *Two-frame interpolation* synthesizes intermediate frames between the cycle extrema [Kim et al., 2024, Guo et al., 2020, Chang et al., 2023, You et al., 2025b]. *Time-conditioned decoders* inject a scalar phase signal into a deformation network [Krebs et al., 2019, Zhang et al., 2025a, Zuler et al., 2025]. *Coordinate-based and per-instance fitting* parameterize motion as a continuous neural field [Yuan et al., 2023, Garzia et al., 2024, Fu et al., 2025, Ye et al., 2023, Zhang et al., 2025b]; mesh-based variants [Meng et al., 2024, Qiao et al., 2025, Ma et al., 2025] model cardiac motion at the mesh-vertex level. Across these forms, the cyclic structure of the heart is learned from data rather than built into the representation: each method derives temporal behavior from per-frame supervision or per-instance optimization rather than from a representation that is cyclic by design.

Spectral motion representations. The closest cardiac precedent for a Fourier representation of motion is the classical, non-deep B-spline–Fourier regularization of Wiputra et al. [2020], which fits Fourier-series coefficients to pair-wise displacement fields as a cyclicity smoothness prior; HARP [Osman et al., 1999] performs spectral-peak isolation in the image domain on tagged MRI rather than on the deformation field. In adjacent domains, *Generative Image Dynamics* [Li et al., 2024] predicts a per-pixel *spectral volume*—Fourier coefficients of long-term oscillatory motion trajectories—via a frequency-coordinated diffusion model, and FB-Diff [You et al., 2025a] extends this idea to respiratory 4D medical imaging with a Fourier-basis-guided diffusion. To our knowledge, no published cardiac deep-learning work predicts Fourier coefficients of per-site deformation trajectories and reconstructs displacement at arbitrary phases by inverse FT. CycleSpectra addresses this gap.

3 Method

3.1 Overview

Let $\mathbf{X}_a$ and $\mathbf{X}_c$ denote two cardiac CT volumes acquired by the same ECG-gated 4D protocol at known cardiac phases $\psi_a, \psi_c \in [0, 1)$ and sampled from opposite portions of the cardiac cycle—one in systole and the other in diastole. We designate $\mathbf{X}_a$ as the anchor frame and $\mathbf{X}_c$ as the context frame. Given these two sparse observations, our goal is to synthesize the cardiac volume at any query phase $\psi_q \in [0, 1)$ by predicting an anchor-to-query deformation field $\boldsymbol{\phi}(\mathbf{x}, \psi_a, \psi_q)$:

$$\hat{\mathbf{X}}_{\psi_q} = \mathcal{W}\big(\mathbf{X}_a,\, \boldsymbol{\phi}(\,\cdot\,, \psi_a, \psi_q)\big), \tag{1}$$

where $\mathcal{W}$ denotes a differentiable warping operator. The two input frames are used to infer a cycle-motion representation, while the query phase is used only to read out the deformation. For queries on the arc closer to the context frame, the roles of the two frames are exchanged (bidirectional read-out, Sec. 3.2).

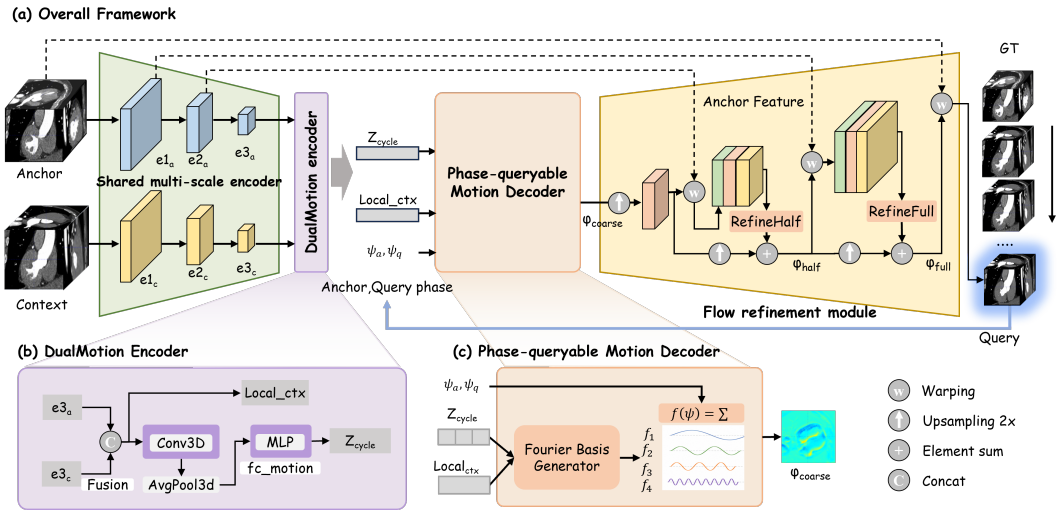

Figure 2: **CycleSpectra overview.** A shared encoder extracts anatomy features from anchor and context volumes; the DualMotion Encoder (b) produces a spatial deformation context and a global cycle code; the phase-queryable decoder (c) predicts Fourier coefficient maps in the latent feature space and analytically synthesizes a coarse flow; two refinement blocks recover the image-scale field.

The core idea of CycleSpectra is to parameterize the cyclic cardiac displacement trajectory as a set of Fourier coefficient maps in a low-dimensional latent space, and to recover the deformation at any query phase through a fixed analytic synthesis layer rather than through a learned phase-conditioned decoder. Time is therefore a coordinate of the analytic read-out rather than a network input, so the representation is periodic and continuous and the cycle structure is decoupled from the phases used for supervision.

The framework is shown in Fig. 2. A shared multi-scale encoder first extracts hierarchical anatomy features from the anchor and context volumes, producing two feature pyramids in a common space. The DualMotion Encoder then compares the deep anchor/context features and outputs two complementary motion descriptors: a local context feature $\mathbf{Local}_{\text{ctx}}$ that preserves voxel-level deformation cues from both observed phases, and a global cycle-motion code $\mathbf{Z}_{\text{cycle}}$ that summarizes the periodic motion pattern implied by the pair. Conditioned on these two descriptors, the phase-queryable motion decoder predicts the Fourier coefficient maps of the displacement trajectory and analytically synthesizes a coarse anchor-to-query flow at the latent scale. Finally, a flow-refinement module progressively upsamples and corrects this coarse field with anchor and context image features, yielding the image-level deformation $\boldsymbol{\phi}(\,\cdot\,, \psi_a, \psi_q)$ used in Eq. (1).

3.2 Frequency-Domain Motion Representation

Latent-space Fourier-coefficient encoding. We parameterize the cyclic cardiac displacement trajectory of each spatial site in a low-dimensional latent feature space rather than at full image resolution. The motion is encoded as K pairs of latent sine and cosine coefficient maps $\{\mathbf{A}_k, \mathbf{B}_k\}_{k=1}^{K}$, where each $\mathbf{A}_k, \mathbf{B}_k \in \mathbb{R}^3$ encodes the three latent displacement components for one Fourier band. The choice of K is small, reflecting the low-frequency nature of cardiac cycle motion.

Analytic phase synthesis. We model the cardiac displacement trajectory at each spatial site $\mathbf{x}$ as a finite Fourier series in cardiac phase $\psi \in [0, 1)$, so the deformation that maps the anchor configuration (at ψ_a) to the queried configuration (at ψ_q) is recovered by the closed-form anchor-relative inverse Fourier transform:

$$\boldsymbol{\phi}(\mathbf{x}, \psi_a, \psi_q) = \sum_{k=1}^{K} \mathbf{A}_k(\mathbf{x}) \big[\sin(2\pi k\psi_q) - \sin(2\pi k\psi_a)\big] + \mathbf{B}_k(\mathbf{x}) \big[\cos(2\pi k\psi_q) - \cos(2\pi k\psi_a)\big]. \quad (2)$$

Eq. (2) is a parameter-free analytic layer: the query phase ψ_q does not enter the network that predicts the coefficient maps, but only the analytic read-out (and the phase-FiLM of the refinement blocks,

Sec. 3.3), and the same coefficient set $\{\mathbf{A}_k, \mathbf{B}_k\}$ describes the displacement trajectory at every phase.

Anchor-relative read-out and context conditioning. Every instance of Eq. (2) is anchor-relative by design: it synthesizes the deformation from a source configuration at reference phase ψ_{ref} to queried configuration ψ_q, with displacement identically zero at $\psi_q{=}\psi_{\text{ref}}$. We use a single source phase per call: adding a second reference phase inside the same read-out would break this zero-at-reference property. We write $\boldsymbol{\phi}(\mathbf{x}, \psi_a, \psi_q)$ for the anchor-sourced call and $\boldsymbol{\phi}(\mathbf{x}, \psi_c, \psi_q)$ for the context-sourced call. The context enters in two stages: *(i) Encoding:* the context *volume* conditions the shared coefficient fields $\{\mathbf{A}_k, \mathbf{B}_k\}$ through appearance correspondence in the DualMotion encoder, consuming paired features $(\mathbf{e}_a^3, \mathbf{e}_c^3)$. *(ii) Read-out:* the context *phase* enters analytically via bidirectional routing to the nearer observed endpoint. The forward arc uses the anchor as source, $\boldsymbol{\phi}(\mathbf{x}, \psi_a, \psi_q)$, and the reverse arc uses the context as source, $\boldsymbol{\phi}(\mathbf{x}, \psi_c, \psi_q)$, entering directly through $\mathbf{s}(\psi_q) - \mathbf{s}(\psi_c)$ where $\mathbf{s}(\psi) = \big(\sin 2\pi k\psi, \cos 2\pi k\psi\big)_{k=1}^{K}$ is the basis vector of Eq. (2). Both directions are read from one shared spectrum, coupling endpoints by construction.

Two operational corollaries. The frequency-domain parameterization yields two important properties. *(i) Arbitrary-phase inference.* Any query phase $\psi_q \in [0, 1)$ becomes a closed-form readout of Eq. (2), with no interpolation between observed frames and no per-case fitting. *(ii) Arbitrary-phase backward supervision.* Any known frame $(\mathbf{X}_{\psi_j}, \psi_j)$ produces a gradient that flows into the same coefficient set $\{\mathbf{A}_k, \mathbf{B}_k\}$ via Eq. (2), so any two-frame full-cycle reconstruction problem reduces to "estimate the per-site coefficient set such that the analytic readout matches every supervised frame at its known phase." We exploit (ii) directly in the two-frame generation regime (Sec. 4.2).

From sparse observations to a full spectrum. It may appear counter-intuitive that two frames suffice to determine a periodic trajectory described by K Fourier components per spatial site. We emphasize that the proposed representation is not estimated frame-by-frame at inference: the coefficient maps $\{\mathbf{A}_k, \mathbf{B}_k\}$ are predicted by a network trained with dense per-phase supervision across the cohort (Sec. 4.2), so the two test-time inputs serve as conditioning rather than as samples of an isolated signal-estimation problem. The network amortizes a population-level prior over cyclic cardiac motion—empirically low-frequency, with low orders ($K = 1$–2) already sufficient (Sec. 4.4)—and the two anchor/context phases inform the prior of where on the cycle the subject currently is. "Sparse" thus refers to test-time conditioning only; the dense training supervision is ablated in Sec. 4.4.

3.3 Network Architecture

Shared multi-scale encoder. A single 3-stage 3D convolutional backbone is applied with shared weights to the anchor volume $\mathbf{X}_a$ and the context volume $\mathbf{X}_c$, producing two feature pyramids $(\mathbf{e}^1, \mathbf{e}^2, \mathbf{e}^3)$ at 1, $1/2$, and $1/4$ of the input grid. Sharing weights across the two inputs guarantees that anchor and context features live in the same space at every scale, so that downstream modules can compare them by simple channel-wise concatenation rather than through a cross-encoder fusion. The encoder receives no phase signal and acts as an anatomy-only descriptor; phase enters only through the analytic read-out of Eq. (2) and the phase-FiLM of the refinement blocks.

DualMotion encoder. The DualMotion encoder (Fig. 2b) turns the paired bottlenecks $(\mathbf{e}_a^3, \mathbf{e}_c^3)$ into the two ingredients required by the analytic synthesis layer in Eq. (2): a spatial site that hosts the per-site coefficient maps, and a global signal that drives them. The two volumes are channel-concatenated and projected by a $1\times1\times1$ fusion convolution into the local-context feature $\mathbf{Local}_{\text{ctx}}$, which preserves spatial-level deformation cues from both observed phases. The same fused volume is then globally average-pooled and mapped by a two-layer MLP into a stack of K frequency tokens $\{\mathbf{z}_k\}_{k=1}^{K}$, collectively denoted $\mathbf{Z}_{\text{cycle}}$. The K tokens are matched one-to-one with the K Fourier bands in Eq. (2) so that every band has its own global control signal. This dual output separates *where* the deformation lives ($\mathbf{Local}_{\text{ctx}}$, dense and spatial) from *how* the cycle behaves ($\mathbf{Z}_{\text{cycle}}$, low-dimensional and global). Because the global pooling is symmetric in $(\mathbf{e}_a^3, \mathbf{e}_c^3)$, $\mathbf{Z}_{\text{cycle}}$ does not encode which frame is the anchor; the anchor identity is reintroduced only at the decoder via the explicit phases ψ_a, ψ_q.

Phase-queryable motion decoder. Given $(\mathbf{e}_a^3, \mathbf{Local}_{\text{ctx}}, \mathbf{Z}_{\text{cycle}}, \psi_a, \psi_q)$, the decoder (Fig. 2c) produces the latent-scale anchor-to-query deformation $\boldsymbol{\phi}^{(3)}(\,\cdot\,, \psi_a, \psi_q)$. We first build the anchor-side

spatial backbone by concatenating $\mathbf{e}_a^3$ with $\mathbf{Local}_{\text{ctx}}$ and projecting back through a $1\times1\times1$ context-fusion convolution; the resulting volume $\tilde{\mathbf{e}}_a^3$ is the substrate on which all K coefficient maps are predicted. The decoder is realized as K parallel stems sharing $\tilde{\mathbf{e}}_a^3$ as input, each FiLM-modulated by its own frequency token $\mathbf{z}_k$ to produce a band-specific hidden volume $\mathbf{h}_k$. A cross-frequency self-attention block then treats $\{\mathbf{h}_k\}_{k=1}^K$ as K tokens at every spatial location and runs multi-head attention along the frequency axis only; this lets neighbouring bands exchange information per spatial site and acts as a spectral regularizer that keeps the bands non-redundant when K is small. A per-band $1\times1\times1$ head then maps each refined $\mathbf{h}_k$ to the per-site cosine and sine coefficient maps $(\mathbf{B}_k(\mathbf{x}), \mathbf{A}_k(\mathbf{x}))$ of Eq. (2). The latent-scale field is then read out in closed form by substituting (ψ_a, ψ_q) into Eq. (2) and summing over the K bands. A useful property of this design is that the K coefficient maps are independent of ψ_q: they are computed once per anchor/context pair and shared across all queries, so the entire decoder is amortized over phases.

Flow refinement module. The decoder predicts $\boldsymbol{\phi}^{(3)}$ at the bottleneck scale; we recover the image-resolution deformation $\boldsymbol{\phi}(\,\cdot\,, \psi_a, \psi_q)$ through two phase-conditioned residual blocks placed at the half- and full-resolution scales, mirroring the encoder strides. At each level the current flow is trilinearly upsampled with per-axis rescaling so that voxel-coordinate displacements remain consistent across scales. Each block is target-aware: it ingests the warped anchor feature, the context-side reference feature at the same scale, and the upsampled flow, and emits a 3-channel correction added to the upsampled flow. A FiLM modulation driven by a sin/cos encoding of $(\psi_a, \psi_q, \Delta\psi)$ injects phase awareness without introducing phase-specific spatial filters, so queries at different ψ_q share all spatial parameters at every scale. The output convolution of each block is zero-initialized, so both refinement blocks act as the identity at training start; this forces the analytic Fourier branch to fit the bulk of the cardiac deformation first, while the convolutional refinement only has to clean up high-frequency residuals near tissue boundaries once the coarse spectrum is in place.

3.4 Training Objective: Endpoint-Closure Consistency

Besides supervising the query frame, we require the predicted query to be able to return to the held-out endpoint. The predicted volume $\hat{\mathbf{X}}_{\psi_q}$ is passed back through the shared encoder and the DualMotion encoder to produce a fresh set of motion latents; the decoder then reads out, via Eq. (2), a flow from ψ_q to the held-out endpoint phase ψ_e, and the resulting warp is matched to the held-out endpoint volume $\mathbf{X}_e$ in image space:

$$\mathcal{L}_{\text{ep}} = \left\| \mathcal{W}\big(\hat{\mathbf{X}}_{\psi_q},\, \hat{\boldsymbol{\phi}}(\,\cdot\,, \psi_q, \psi_e)\big) - \mathbf{X}_e \right\|_1. \tag{3}$$

This gives an additional gradient path through Eq. (2) from the model's own intermediate prediction and ties both observed endpoints to the same spectrum; removing it lowers LV/RV curve correlation from 0.862 to 0.845 (Tab. A11).

4 Experiments

4.1 Datasets

In-house cohort. We acquire a phase-resolved 4D cardiac CT cohort of 36 subjects from our clinical partner, with thin-slice resolution (1.6 mm) and 6–24 phases per subject. The cohort contains two subjects with reduced ejection fraction (GT LVEF $< 40\%$), one of which lies in the test set; it carries no validated labels for arrhythmia, regional wall-motion abnormality, congenital disease, or post-operative anatomy. The 85 opposite pairs are nested within the same 7 test subjects: they broaden endpoint coverage, not the number of independent anatomies. Subjects are split into 22 training, 7 validation, and 7 testing, stratified by frame count and LVEF; held-out test subjects are never used for checkpoint selection.

Data governance. Scans were fully de-identified before transfer (removal of patient names, MRNs, dates, and direct DICOM metadata). Under data-use agreements, raw CT cannot be shared.

Out-of-distribution evaluation. We additionally evaluate on an external 4D-CT cohort and a cross-modality 4D-cMR set without retraining; full results are deferred to App. A.2.

4.2 Implementation details

Network configuration. The shared image encoder uses base channel width $C{=}24$, so the bottleneck $\mathbf{e}^3$ has $4C{=}96$ channels at $1/4$ of the input grid. The DualMotion encoder produces $K{=}4$ frequency tokens of dimension $d_m{=}16$, paired one-to-one with the K Fourier bands. The phase-queryable decoder uses hidden width $H_d{=}64$ inside its K FiLM stems and 4-head attention along the frequency axis. The two target-aware flow-refinement blocks have hidden widths 32 (half-resolution, operating on $\mathbf{e}_a^2/\mathbf{e}_c^2$) and 16 (full-resolution, $\mathbf{e}_a^1/\mathbf{e}_c^1$), both with zero-initialized output convolutions. The latent-scale displacement field is clipped to ± 30 voxels per axis before the refinement stack.

Training. Training uses four NVIDIA RTX 3090 GPUs and takes about 4 hours for 100 epochs; the reported checkpoint is the epoch with the best validation PSNR (epoch 81). We train with AdamW ($\beta_1{=}0.9$, $\beta_2{=}0.999$, weight decay 10^{-5}) at base learning rate $5{\times}10^{-5}$, with a 3-epoch linear warmup followed by cosine decay to 10^{-6} over 100 epochs. Global gradient clipping is 1.0 and mixed-precision training is enabled.

Data sampling. Every volume is heart-centered-cropped, downsampled by a factor of 2 to a fixed $64{\times}128{\times}128$ grid, and intensity-normalized to $[0, 1]$. The only training augmentation is a random heart-crop jitter of up to 10 voxels along each axis. A training sample is a triplet $(\mathbf{X}_a, \mathbf{X}_c, \{\mathbf{X}_{\psi_j}\}_j)$, where the anchor/context endpoints $(\mathbf{X}_a, \mathbf{X}_c)$ form an *opposite-portion random pair*: one frame is drawn from the systolic arc and the other from the diastolic arc of the cardiac cycle, with a minimum frame-index gap of 3. The supervision set $\{\mathbf{X}_{\psi_j}\}_j$ sweeps all remaining cardiac phases of that subject (one full cycle per epoch per subject), so every training step receives anchor-relative supervision at every available phase, directly exercising the arbitrary-phase supervision property of Eq. (2).

Loss. The total training loss is

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\ell_1} + \lambda_{\text{rec}}\lambda_{\text{ssim}}\mathcal{L}_{\text{SSIM}} + \lambda_{\text{fg}}\mathcal{L}_{\text{fg}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{fold}}\mathcal{L}_{\text{fold}} + \lambda_{\text{ep}}\mathcal{L}_{\text{ep}}, \tag{4}$$

with weights $\lambda_{\text{rec}} = 1.0$ (masked ℓ_1 on the warped query image, App. A.1.2), $\lambda_{\text{ssim}} = 0.2$, $\lambda_{\text{fg}} = 1.0$ (foreground-mask reweighting on the heart ROI), $\lambda_{\text{smooth}} = 0.05$ (flow first-order smoothness), $\lambda_{\text{fold}} = 0.05$ (Jacobian folding penalty), and $\lambda_{\text{ep}} = 0.1$ for the endpoint-closure consistency term (Sec. 3.4, Eq. (3)).

Evaluation protocols. We evaluate two-frame full-cycle generation under two protocols on the same 7 held-out test subjects. The *canonical ED/ES* protocol fixes one ED/ES endpoint pair per subject (7 pairs in total), matching the common endpoint-supervised setting in prior work. The *opposite-pair* protocol enumerates every endpoint pair whose two phases lie on opposite cyclic arcs, with cyclic phase distance in $[0.35, 0.65]$ and frame-index gap > 3 (85 pairs in total across the same 7 subjects). Both protocols evaluate every non-input query frame (1,038 frames for the opposite-pair bank). Metrics are aggregated frame $\rightarrow$ pair $\rightarrow$ subject $\rightarrow$ cohort: the seven subjects are weighted equally, "$\pm$" denotes the population SD over the seven subjects, and intervals are 95% percentile bootstrap intervals over subjects (10,000 resamples; App. A.7). The 85 opposite pairs are nested within the seven subjects and are never treated as independent test cases. For the functional metrics, the reference volume–time curve uses a frozen chamber segmentation of every native 4D-CT phase; the Curve metric is the LV/RV average of the Pearson correlation over non-input phases. The predicted chamber mask follows each method's native output (Tab. A4). The model checkpoint is identical across the two protocols.

Inference. At test time, the encoder and DualMotion encoder run once per input pair to produce $K{=}4$ pairs of latent coefficient maps; query phases ψ_q are then synthesized analytically via Eq. (2) followed by the two refinement blocks, sharing all spatial parameters across queries. The full T-frame trajectory is produced in one shot rather than autoregressively. Inference runs on a single RTX 3090.

4.3 Results

Baselines and fairness. All methods use *matched inference information and evaluation protocol, with method-specific training objectives*: each receives only the same two observed CT volumes of a held-out subject at $64{\times}128{\times}128$, predicts the same query phases, never accesses query ground truth, and is scored identically. Temporal interfaces, objectives, and training-pair exposure differ by design (App. A.5, Tab. A4). Dyna3DGR is not a frozen model but is fitted from scratch on every

Table 1: **Two-frame full-cycle generation results** (mean ± population SD over 7 held-out subjects; pairs averaged within subject). Image fidelity is measured over full volume and heart ROI; anatomical/motion metrics exclude input frames. Curve correlation is the LV/RV average; LVEF error is the absolute difference in ejection fraction (percentage points). Both blocks share one checkpoint (Sec. 4.2). Chamber masks follow native outputs (Tab. A4): CycleSpectra/Dyna3DGR warp the source label; UVI-Net/Identity hold it fixed; T0 linear-flow warps it with the scaled displacement field; FusionNet, TMKD, Mo-Diff are re-segmented with a frozen nnU-Net [Isensee et al., 2021]; TTT4D mask provenance could not be verified. A common-segmenter control reproduces the ranking (Tab. A6–A7). Bold marks the best nominal mean, not statistical significance (paired intervals in Tab. A10). Mo-Diff is a diffusion model whose data requirements exceed our cohort; its chamber Dice is near zero, so its distance metrics (valid surfaces only) should not be read as favorable geometry.

Method	PSNR↑	SSIM↑	ROI PSNR↑	ROI SSIM↑	Dice avg↑	HD95↓	Curve corr↑	SV err.↓	LVEF err.↓
Canonical ED/ES									
Identity	20.31±0.91	0.752±0.034	18.56±0.87	0.660±0.025	0.751±0.066	11.04±6.07	-0.093±0.235	38.12±23.91	7.32±6.04
T0 linear-flow	21.70±0.91	0.790±0.030	20.03±0.86	0.715±0.022	0.776±0.059	10.64±6.26	0.158±0.282	39.58±23.45	7.65±6.18
UVI-Net (CVPR'24)	**24.18**±0.92	**0.845**±0.019	**22.29**±0.91	**0.776**±0.015	0.751±0.066	11.04±6.07	-0.093±0.235	38.12±23.91	7.32±6.04
TTT4D	21.06±1.26	0.789±0.028	19.39±1.21	0.704±0.025	0.719±0.062	12.14±6.09	-0.031±0.285	49.85±26.99	**5.52**±4.79
TMKD (MICCAI'25)	22.88±0.82	0.810±0.009	21.00±0.79	0.737±0.008	0.778±0.059	10.54±4.06	0.248±0.222	42.15±13.14	6.37±4.93
FusionNet (MICCAI'23)	21.24±0.93	0.787±0.027	19.71±1.00	0.742±0.023	0.778±0.061	13.32±8.26	**0.781**±0.181	41.44±39.08	9.01±13.49
Mo-Diff (MICCAI'25)	15.44±2.06	0.495±0.080	14.24±2.08	0.443±0.078	0.097±0.062	7.46±1.85*	0.143±0.233*	156.52±50.21	43.00±11.35†
Dyna3DGR (MICCAI'25)	19.95±0.40	0.643±0.013	19.66±0.80	0.640±0.009	0.713±0.061	11.96±5.97	-0.060±0.244	**37.57**±23.62	6.90±4.80
Ours	22.60±0.82	0.807±0.027	21.09±0.76	0.745±0.022	**0.800**±0.061	**9.83**±6.39	0.632±0.241	38.03±24.36	7.80±8.90
Opposite-pair									
Identity	21.30±0.77	0.777±0.032	19.54±0.69	0.695±0.025	0.785±0.047	8.87±2.65	0.102±0.188	57.39±6.93	24.41±5.90
T0 linear-flow	22.19±0.74	0.800±0.031	20.49±0.63	0.728±0.028	0.796±0.047	8.64±2.69	0.340±0.172	55.06±8.03	21.11±6.80
UVI-Net (CVPR'24)	**23.55**±0.75	**0.835**±0.026	21.62±0.64	0.760±0.025	0.785±0.047	8.87±2.65	0.102±0.188	57.39±6.93	24.41±5.90
TTT4D	20.61±1.23	0.778±0.034	18.91±1.17	0.688±0.031	0.740±0.048	10.29±2.71	0.185±0.153	46.83±7.58	11.40±2.54
TMKD (MICCAI'25)	23.01±0.75	0.812±0.012	21.09±0.64	0.741±0.015	0.778±0.058	10.62±3.83	0.187±0.166	64.53±7.41	24.47±5.96
FusionNet (MICCAI'23)	21.21±1.20	0.785±0.028	19.70±1.27	0.739±0.027	0.779±0.061	13.60±7.69	0.853±0.128	46.03±36.64	9.41±11.26
Mo-Diff (MICCAI'25)	15.76±0.86	0.500±0.031	14.59±0.92	0.452±0.028	0.061±0.014	13.13±9.39	0.196±0.224	97.56±34.50	43.00±11.35†
Dyna3DGR (MICCAI'25)	19.82±0.33	0.640±0.011	19.47±0.63	0.637±0.009	0.739±0.048	9.83±2.47	0.113±0.190	55.80±7.02	23.29±5.86
Ours	23.13±0.68	0.820±0.026	**21.67**±0.65	**0.763**±0.023	**0.821**±0.050	**7.94**±2.91	**0.862**±0.075	**44.12**±11.89	**8.00**±6.37

*Mo-Diff canonical HD95 and Curve correlation are computed over $n = 5$ valid subjects due to chamber segmentation failure on sub4 and sub8, and are excluded from bolding. †Mo-Diff produces at least one empty predicted LV mask in every evaluated cycle, causing predicted LVEF to saturate at 100% in both protocols; the identical LVEF errors reflect this segmentation collapse rather than duplicated data.

test pair (20,000 iterations, 12,288 nodes), and our FusionNet is a two-endpoint 3D-CT adaptation of its multi-path design, not an unchanged run of the original five-frame LV-shape model. Prior interpolation methods are designed around ED/ES inputs; recovering the cycle from arbitrary phase pairs is the setting studied here.

We compare CycleSpectra against interpolation, registration-derived, and generative two-frame baselines on the same 7 held-out subjects (Tab. 1); the two-frame regime matches the input setup of all baselines, and our framework also supports multi-frame extensions (App. A.4). The opposite-pair protocol, unlike canonical ED/ES, tests whether a method has learned a reusable cyclic motion representation rather than a privileged ED-to-ES interpolation rule.

In-domain comparison. The canonical ED/ES block in Tab. 1 shows why image-only metrics are insufficient for this task: interpolation and alpha-blending baselines obtain competitive full-volume PSNR because much of the CT volume is static background, yet their motion-sensitive metrics remain weak. The heart-ROI and anatomical columns reduce this background dilution and reveal the actual quality of the reconstructed motion. Under the opposite-pair protocol, where the inputs are no longer the privileged ED/ES extrema, CycleSpectra has the best nominal mean on anatomical agreement (Dice 0.821, HD95 7.94 mm) and on functional consistency (LV/RV curve correlation 0.862, LVEF error 8.00%) among methods with comparable image fidelity. Because only seven subjects are tested, we report subject-level 95% bootstrap intervals (Tab. A8) and state superiority only where the paired interval of the method difference excludes zero (Tab. A10). On this criterion CycleSpectra is better than every baseline on Dice, and better than UVI-Net, TMKD, and Dyna3DGR on curve correlation and LVEF error; its edge over FusionNet (curve correlation 0.862 vs. 0.853, paired difference $+0.008$ $[-0.072, +0.100]$; LVEF error 8.00 vs. 9.41%, -1.41 pp $[-6.28, +2.67]$) and over TTT4D (LVEF error 11.40%, -3.40 pp $[-8.20, +2.91]$) is a nominal-mean ordering whose interval includes zero. Dyna3DGR refits per pair and has no shared cycle prior,

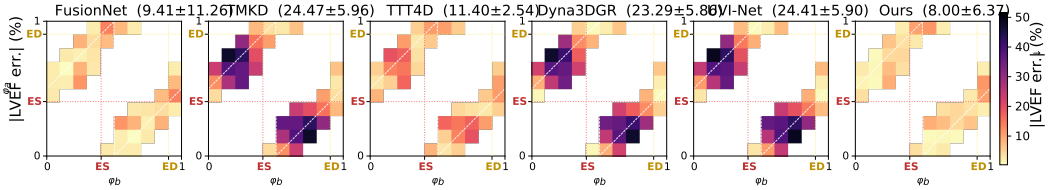

Figure 3: **Robustness to opposite-pair anchor choice.** Per-method heatmaps of |LVEF err.| over the (endpoint a, endpoint b) phase grid on the test split (85 pairs / 7 subjects). Dotted lines mark cohort-mean canonical ED ($\bar{\psi}_{\mathrm{ED}} \approx 0.90$) and ES ($\bar{\psi}_{\mathrm{ES}} \approx 0.40$). Compared with baselines whose error spikes away from the canonical intersection, our method keeps LVEF error low across most of the opposite-pair plane (the exception is the single low-EF subject, Sec. 5).

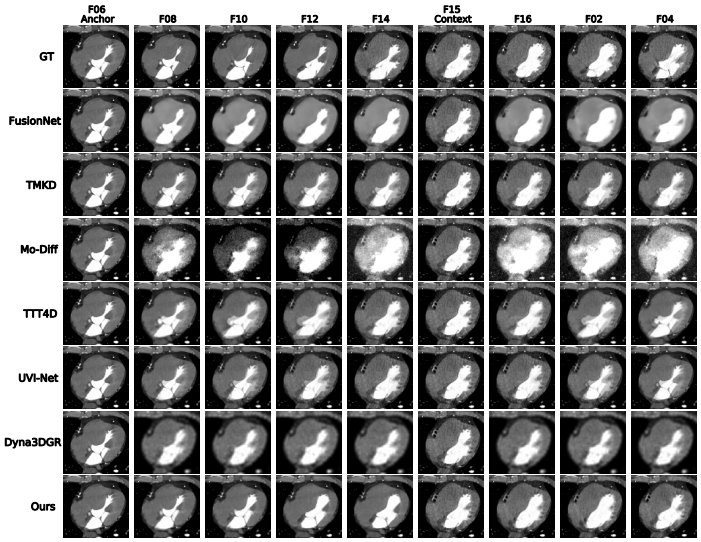

Figure 4: **Qualitative opposite-pair full-cycle generation.** Anchor F06 and context F15 are used as the two-frame input. Rows: methods. Columns: cardiac phases.

so its motion-sensitive metrics collapse on both blocks (canonical curve corr -0.060; opposite curve corr 0.113, LVEF err 23.29%) despite moderate Dice. Under canonical ED/ES, CycleSpectra is not uniformly best: FusionNet has higher curve correlation (0.781 vs. 0.632) and Dyna3DGR marginally lower SV error (37.57 vs. 38.03%), consistent with ED/ES favoring interpolation methods.

Robustness across the opposite-pair plane. Fig. 3 shows per-method LVEF error over the endpoint-phase plane (85 pairs). Interpolating methods (UVI-Net, TMKD) are accurate only near the canonical ED/ES intersection; FusionNet and TTT4D drift more mildly. CycleSpectra keeps LVEF error low across most of the plane (exception: the single low-EF subject, Sec. 5), indicating a reusable cycle trajectory rather than pair-specific interpolation. Per-pair distributions and further heatmaps are in App. A.3.

Local motion fidelity and common-segmenter control. Regional strain and dense displacement cannot be validated on this cohort (no tagged MRI, landmarks, or dense deformation ground truth), so we report two proxies (App. A.6). CycleSpectra is near-incompressible and essentially fold-free (0.009% folded voxels, volume distortion 0.082 vs. 0.200 for Dyna3DGR; Tab. A5). Re-scoring all seven methods with a single frozen no-TTA nnU-Net reproduces the ranking (cardiac-5 macro Dice 0.811 vs. $\leq$0.791; HD95 9.03 vs. $\geq$9.36 mm; best on LV and myocardium, while a few baselines lead on isolated right-heart distances; Tab. A6–A7).

Qualitative comparison. Fig. 4 shows an opposite-pair sequence from anchor F06 and context F15. Two-frame interpolation baselines produce blurred or quasi-static intermediate frames once inputs leave the canonical extrema, while CycleSpectra preserves chamber boundaries and recovers the contraction–relaxation motion across the cycle. Transfer to an external 4D-CT cohort and a cross-modality 4D-cMR set without retraining is reported in App. A.2.

Table 2: **Ablation on spectral motion representation** (opposite-pair bank, 7-subject mean ± population SD). Upper: SinPE flow decoder vs. spectral. Lower: Fourier order K.

Study	Variant	Params (M)↓	PSNR↑	ROI PSNR↑	Dice avg↑	HD95↓	LV/RV curve corr↑	SV err. (%)↓	LVEF err. (%)↓
Representation	SinPE flow decoder	1.004	22.76±0.71	21.17±0.69	0.812±0.048	8.28±2.78	0.718±0.099	55.28±5.56	19.69±7.44
Representation	**Ours** (spectral, K=4)	1.857	23.13±0.68	21.67±0.65	0.821±0.050	7.94±2.91	0.862±0.075	44.12±11.90	8.00±6.37
Fourier order	$K = 1$	1.013	23.03±0.66	21.48±0.66	0.819±0.049	8.01±2.91	0.869±0.072	41.80±10.87	7.52±6.96
Fourier order	$K = 2$	1.294	23.16±0.67	21.68±0.63	0.823±0.050	7.89±2.93	0.870±0.064	42.44±10.04	7.32±7.50
Fourier order	$K = 4$ (default)	1.857	23.13±0.68	21.67±0.65	0.821±0.050	7.94±2.91	0.862±0.075	44.12±11.90	8.00±6.37
Fourier order	$K = 6$	2.420	23.02±0.67	21.60±0.64	0.821±0.051	7.92±2.96	0.868±0.068	42.34±11.05	8.03±6.68

4.4 Ablation Study and Representation Analysis

Spectral representation vs. a sinusoidal phase-conditioned flow decoder. To isolate the frequency-domain representation, we replace the Fourier branch with a *sinusoidal phase-conditioned flow decoder* (SinPE flow decoder): a free-form flow head conditioned on phase sine/cosine encodings with identical encoder, refinement, and training. As shown in Tab. 2 (upper), SinPE attains similar image and anatomical scores (PSNR 22.76 vs. 23.13 dB, SSIM 0.814 vs. 0.820, Dice 0.812 vs. 0.821), but motion-sensitive metrics degrade sharply: curve correlation falls from 0.862 to 0.718 (-0.144) and LVEF error rises from 8.00 to 19.69% ($\approx$2.5$\times$). This suggests that sine/cosine phase encodings alone do not yield cyclic whole-cycle motion, whereas a shared per-site spectrum with an analytic read-out does.

Fourier order. Varying K in Tab. 2 (lower) shows that motion-sensitive metrics are stable across $K \in \{1, 2, 4, 6\}$ (curve corr 0.862–0.870, LVEF error 7.32–8.03%, SV error 41.8–44.1%), consistent with cardiac motion being low-frequency. We retain K=4 as a conservative default, while K=2 is nominally best on curve correlation and Dice with differences well within subject variance.

Loss design. Tab. A11 (App. A.8) ablates each loss term. Removing foreground weighting hurts most (curve corr 0.862 → 0.824, SV error 44.12 → 47.33%, LVEF error 8.00 → 10.23%); removing endpoint closure degrades consistently (curve corr → 0.845; PSNR -0.12 dB, HD95 $+0.08$ mm). SSIM and smoothness/folding terms leave point-estimate motion metrics unchanged but are kept for image quality and deformation regularity.

Sparse training supervision. "Sparse" refers to test-time conditioning, not training supervision. Retraining an otherwise identical model on only ED, ES, and one mid-arc query per subject (66 of 318 training phases; no teacher; App. A.9) meaningfully lowers image fidelity (PSNR 23.134 → 20.912 dB, paired difference -2.223 [-2.467, -1.979]) and degrades motion metrics most (curve corr 0.862 → 0.432, -0.430 [-0.523, -0.336]; LVEF error 7.995 → 14.560%, interval includes zero): dense phase supervision is an important ingredient of the learned motion prior.

5 Conclusion and Limitations

CycleSpectra parameterizes cyclic cardiac displacement as per-site Fourier coefficient maps in a low-dimensional latent space, decoded at any phase via a fixed analytic synthesis layer. Periodicity and continuity hold by design. On a thin-slice 4D cardiac CT cohort, CycleSpectra achieves the best mean anatomical and functional metrics under non-canonical opposite pairs with competitive image fidelity, and remains competitive under canonical ED/ES. Our framework also admits multi-frame extensions with consistent gains across in-house and cross-modality cohorts (App. A.4).

Limitations. (i) *Cohort and pathology:* 36 subjects (7 test subjects with nested pairs, so several comparisons, notably vs. FusionNet, are nominal orderings with overlapping intervals; raw CT unshareable), without validated labels for wall-motion abnormalities or arrhythmia; the single low-EF subject (GT LVEF 32.3%) is our clear failure case (LVEF error 22.4% opposite-pair, 29.3% canonical). (ii) *Supervision:* Test-time input is sparse (two frames), but training uses dense per-phase supervision, so the learned cohort-level motion prior depends on densely sampled training phases (Sec. 4.4). (iii) *Stationary cycle:* One representative cycle is modeled as periodic; arrhythmia departs from this. (iv) *Evaluation and scope:* Functional metrics depend on chamber-label provenance (controlled in App. A.6), regional strain is unvalidated, and reconstruction is deterministic without uncertainty quantification (App. A.10).

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China under Grant 62271465, the National Key R&D Program of China under Grant 2025YFC3408300, the Basic Research Program of Jiangsu (No. BK20255001), the Suzhou Basic Research Program under Grant SYG202338, and the Sanming Project of Medicine in Shenzhen under Grant SZSM202411033. We thank our clinical partner and clinicians for the data. **Competing interests:** The authors declare no competing interests.

References

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

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

Chujie Chang, Shoko Miyauchi, Ken'ichi Morooka, Ryo Kurazume, and Oscar Martinez Mozos. FusionNet: A frame interpolation network for 4D heart models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–44. Springer, 2023.

Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. TransMorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, 2022.

Zeyuan Chen, Yuanjie Zheng, and James C Gee. TransMatch: a transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration. *IEEE Transactions on Medical Imaging*, 43(1):15–27, 2024.

Xueming Fu, Pei Wu, Yingtai Li, Xin Luo, Zihang Jiang, Junhao Mei, Jian Lu, Gao-Jun Teng, and Shaohua Kevin Zhou. Dyna3DGR: 4D cardiac motion tracking with dynamic 3D gaussian representation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 164–174. Springer, 2025.

Simone Garzia, Patryk Rygiel, Sven Dummer, Filippo Cademartiri, Simona Celi, and Jelmer M Wolterink. Neural fields for continuous periodic motion estimation in 4D cardiovascular imaging. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 378–389. Springer, 2024.

Yuyu Guo, Lei Bi, Euijoon Ahn, Dagan Feng, Qian Wang, and Jinman Kim. A spatiotemporal volumetric interpolation network for 4D dynamic medical image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4726–4735, 2020.

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

JungEun Kim, Hangyul Yoon, Geondo Park, Kyungsu Kim, and Eunho Yang. Data-efficient unsupervised interpolation without any intermediate frame for 4D medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11353–11364, 2024.

Julian Krebs, Hervé Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE Transactions on Medical Imaging*, 38 (9):2165–2176, 2019.

Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24142–24153, 2024.

Qiang Ma, Qingjie Meng, Mengyun Qiao, Paul M Matthews, Declan P O'Regan, and Wenjia Bai. CardiacFlow: 3D+t four-chamber cardiac shape completion and generation via flow matching. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 89–99. Springer, 2025.

Qingjie Meng, Wenjia Bai, Declan P O'Regan, and Daniel Rueckert. DeepMesh: mesh-based cardiac motion tracking using deep learning. *IEEE Transactions on Medical Imaging*, 43(4):1489–1500, 2024.

Manuel A Morales, Maaike Van den Boomen, Christopher Nguyen, Jayashree Kalpathy-Cramer, Bruce R Rosen, Collin M Stultz, David Izquierdo-Garcia, and Ciprian Catana. DeepStrain: a deep learning workflow for the automated characterization of cardiac mechanics. *Frontiers in Cardiovascular Medicine*, 8:730316, 2021.

Nael F Osman, William S Kerwin, Elliot R McVeigh, and Jerry L Prince. Cardiac motion tracking using CINE harmonic phase (HARP) magnetic resonance imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(6):1048–1060, 1999.

Mengyun Qiao, Kathryn A McGurk, Shuo Wang, Paul M Matthews, Declan P O'Regan, and Wenjia Bai. A personalized time-resolved 3D mesh generative model for unveiling normal heart dynamics. *Nature Machine Intelligence*, 7(5):800–811, 2025.

Chen Qin, Shuo Wang, Chen Chen, Wenjia Bai, and Daniel Rueckert. Generative myocardial motion tracking via latent space exploration with biomechanics-informed prior. *Medical Image Analysis*, 83:102682, 2023.

Matthew Sinclair, Andreas Schuh, Karl Hahn, Kersten Petersen, Ying Bai, James Batten, Michiel Schaap, and Ben Glocker. Atlas-ISTN: joint segmentation, registration and atlas construction with image-and-spatial transformer networks. *Medical Image Analysis*, 78:102383, 2022.

Haiqiao Wang, Zhuoyuan Wang, Dong Ni, and Yi Wang. ModeTv2: GPU-accelerated motion decomposition transformer for pairwise optimization in medical image registration. *Medical Image Analysis*, page 103862, 2025.

Hadi Wiputra, Wei Xuan Chan, Yoke Yin Foo, Sheldon Ho, and Choon Hwai Yap. Cardiac motion estimation from medical images: a regularisation framework applied on pairwise image registration displacement fields. *Scientific reports*, 10(1):18510, 2020.

Jiewen Yang, Yiqun Lin, Bin Pu, and Xiaomeng Li. Bidirectional recurrence for cardiac motion tracking with gaussian process latent coding. *Advances in Neural Information Processing Systems*, 37:34800–34823, 2024.

Meng Ye, Dong Yang, Mikael Kanski, Leon Axel, and Dimitris Metaxas. Neural deformable models for 3D bi-ventricular heart shape reconstruction and modeling from 2D sparse cardiac magnetic resonance imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Xin You, Runze Yang, Chuyan Zhang, Zhongliang Jiang, Jie Yang, and Nassir Navab. FB-Diff: Fourier basis-guided diffusion for temporal interpolation of 4D medical imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28010–28020, 2025a.

Xin You, Minghui Zhang, Hanxiao Zhang, Jie Yang, and Nassir Navab. Temporal differential fields for 4D motion modeling via image-to-video synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 606–616. Springer, 2025b.

Xiaohan Yuan, Cong Liu, and Yangang Wang. 4D myocardium reconstruction with decoupled motion and shape model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21252–21262, 2023.

Jiaju Zhang, Danni Ai, Zhikun Gan, Tianyu Fu, Jingfan Fan, Hong Song, Deqiang Xiao, and Jian Yang. Temporal modulated multi-scale deformation fusion via knowledge distillation for 4D medical image interpolation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 551–561. Springer, 2025a.

Qikang Zhang, Yingjie Lei, Zihao Zheng, Ziyang Chen, and Zhonghao Xie. Test time training for 4D medical image interpolation. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2025b.

Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, 58: 101537, 2019.

Shahar Zuler, Gal Lifshitz, Hadar Averbuch-Elor, and Dan Raviv. Systole-conditioned generative cardiac motion. *arXiv preprint arXiv:2507.15894*, 2025.

Appendix

A.1 Implementation Details

This section provides extended implementation details that are summarized in the main paper.

A.1.1 Network Architecture

Shared multi-scale encoder. The encoder is a 3-stage 3D convolutional backbone with shared weights between the anchor and context inputs. Each stage is a ConvBlock3D — two $3 \times 3 \times 3$ convolutions with GroupNorm and SiLU activations. Consecutive stages are connected by a stride-2 DownBlock3D that halves every spatial axis and doubles the channel count. With base channel width C=24, this produces feature pyramids $\mathbf{e}^1 \in \mathbb{R}^{C \times D \times H \times W}$, $\mathbf{e}^2 \in \mathbb{R}^{2C \times \frac{D}{2} \times \frac{H}{2} \times \frac{W}{2}}$, $\mathbf{e}^3 \in \mathbb{R}^{4C \times \frac{D}{4} \times \frac{H}{4} \times \frac{W}{4}}$.

DualMotion encoder. The paired bottlenecks $(\mathbf{e}_a^3, \mathbf{e}_c^3)$ are channel-concatenated and projected by a $1 \times 1 \times 1$ fusion convolution with GroupNorm-SiLU back to $4C$=96 channels, yielding the local-context feature $\mathbf{Local}_{\text{ctx}}$. The same fused volume is globally average-pooled to a vector $\mathbf{v} \in \mathbb{R}^{4C}$, then mapped by a two-layer MLP (hidden width 128) to a stack of K=4 frequency tokens $\{\mathbf{z}_k\}_{k=1}^K$ of dimension d_m=16, collectively denoted $\mathbf{Z}_{\text{cycle}}$.

Phase-queryable motion decoder. We first build the anchor-side spatial backbone $\tilde{\mathbf{e}}_a^3$ by concatenating $\mathbf{e}_a^3$ with $\mathbf{Local}_{\text{ctx}}$ and projecting back through a 1×1×1 context-fusion conv (GroupNorm-SiLU). The decoder is realized as K parallel stems sharing $\tilde{\mathbf{e}}_a^3$, each FiLM-modulated by its own frequency token $\mathbf{z}_k$ via a linear projection that yields per-channel scale/shift inside a two-layer GroupNorm-SiLU 3D conv block of hidden width H_d=64. The cross-frequency self-attention block runs 4-head multi-head attention along the frequency axis only — treating $\{\mathbf{h}_k\}_{k=1}^K$ as K tokens at every spatial location, this lets neighbouring bands exchange information per spatial site and acts as a spectral regularizer that keeps the bands non-redundant when K is small. Each band's $1 \times 1 \times 1$ output head maps $\mathbf{h}_k$ to 6 channels, which are split into the per-site cosine and sine coefficient maps $(\mathbf{B}_k(\mathbf{x}), \mathbf{A}_k(\mathbf{x})) \in \mathbb{R}^3 \times \mathbb{R}^3$. The latent-scale displacement field is magnitude-clipped to ± 30 voxels per axis before refinement.

Flow refinement module. At each of two refinement levels (half-resolution and full-resolution), the current flow is trilinearly upsampled with per-axis rescaling so that voxel-coordinate displacements remain consistent across scales. Each block ingests the warped anchor feature $\mathcal{W}(\mathbf{e}_a^\ell, \boldsymbol{\phi}^\uparrow)$, the context-side reference feature $\mathbf{e}_c^\ell$, and the upsampled flow, along the channel axis. The block consists of a 3×3×3 conv with GroupNorm-SiLU, a FiLM modulation driven by a 7-d sin/cos phase encoding $[\sin 2\pi\psi_a, \cos 2\pi\psi_a, \sin 2\pi\psi_q, \cos 2\pi\psi_q, \sin 2\pi\Delta\psi, \cos 2\pi\Delta\psi, \Delta\psi]$, a second $3 \times 3 \times 3$ conv with GroupNorm-SiLU, and a zero-initialized $1 \times 1 \times 1$ output that emits a 3-channel correction added to the upsampled flow. The half-resolution block uses $\mathbf{e}_a^2, \mathbf{e}_c^2$ ($2C$=48 channels each, hidden width 32) and the full-resolution block uses $\mathbf{e}_a^1, \mathbf{e}_c^1$ (C=24 channels each, hidden width 16).

A.1.2 Loss Term Definitions

The total training loss in Eq. (4) contains six terms; we define each below. Let $\hat{\mathbf{X}}_{\psi_q}$ and $\mathbf{X}_{\psi_q}$ denote the predicted and ground-truth volumes at query phase ψ_q, $\boldsymbol{\phi}$ the predicted deformation field, and $\mathbf{M}_{\text{fg}}$ the binary heart-ROI foreground mask.

Reconstruction terms. $\mathcal{L}_{\ell_1}$ is the masked ℓ_1 reconstruction loss between the predicted volume $\hat{X}_{\psi_q}$ and the ground-truth volume X_{ψ_q} at the query phase ψ_q, i.e. the anchor-to-query photometric residual:

$$\mathcal{L}_{\ell_1} \;=\; \mathcal{L}_{a \to q} \;=\; \big\| X_{\psi_q} \;-\; \mathcal{W}\big(X_a,\, \phi(\mathbf{x}, \psi_a, \psi_q)\big) \big\|_1, \tag{5}$$

where $\mathcal{W}$ is the differentiable warping operator of Eq. (1) and $\phi(\mathbf{x}, \psi_a, \psi_q)$ is the anchor-to-query deformation field synthesized from the Fourier coefficient maps via Eq. (2). $\mathcal{L}_{\text{SSIM}}$ is the standard SSIM loss between $\hat{X}_{\psi_q}$ and X_{ψ_q} within the heart-centered crop, and $\mathcal{L}_{\text{fg}}$ reweights the reconstruction loss to give higher weight to voxels inside the heart-ROI foreground mask M_{fg}.

Regularization terms. $\mathcal{L}_{\text{smooth}}$ is the standard first-order smoothness penalty on $\boldsymbol{\phi}$, computed as the mean squared norm of spatial gradients of the displacement field. $\mathcal{L}_{\text{fold}}$ penalizes Jacobian fold-

ing: voxels where $\det(\mathbf{J}_{\boldsymbol{\phi}}) \leq 0$ contribute a penalty proportional to the magnitude of the negative determinant.

Endpoint-closure consistency. (see Sec. 3.4) $\mathcal{L}_{ep}$ enforces consistency between the predicted query frame and the held-out endpoint X_e. Concretely, the predicted volume $\hat{X}_{\psi_q}$ is passed back through the encoder and the DualMotion encoder to produce a fresh set of motion latents; the decoder then synthesizes a flow $\hat{\boldsymbol{\phi}}(\cdot, \psi_q, \psi_e)$ from ψ_q to the held-out endpoint phase ψ_e, and the resulting warp of $\hat{X}_{\psi_q}$ is matched to X_e in image space as in Eq. (3). This term provides an additional gradient pathway through Eq. (2) from the model's own intermediate prediction back to a ground-truth endpoint, complementing the dense per-phase supervision described in the main paper.

A.1.3 Data Preprocessing

All cardiac CT volumes are heart-centered-cropped, downsampled by a factor of 2 to a fixed $64 \times 128 \times 128$ grid, and intensity-normalized to $[0, 1]$. Predicted intensity volumes are clamped to $[0, 1]$ before being saved. Reference chamber labels (LV, RV, LA, RA) for every native 4D-CT phase come from a frozen in-house fine-tuned nnU-Net [Isensee et al., 2021]. Predicted query masks follow each method's native output (Tab. A4): CycleSpectra and Dyna3DGR warp the source-endpoint label with the predicted flow (nearest-neighbour); UVI-Net holds it fixed; FusionNet, TMKD, and Mo-Diff are re-segmented from the generated CT with the same frozen nnU-Net; the TTT4D mask provenance could not be verified. The common-segmenter control (App. A.6) applies one frozen no-TTA nnU-Net to every generated CT.

A.2 Out-of-Distribution Evaluation

We additionally evaluate two-frame full-cycle generation on two external cohorts without retraining the model. The single trained checkpoint used for all main-paper experiments is applied as-is.

OOD-4DCT. We use 4 publicly available cardiac CT cases (10 phases each, 1 mm slice thickness) with 59 fixed opposite-phase endpoint pairs. This setting tests scanner and acquisition shift while keeping the modality (CT) consistent with our training cohort. Tab. A1 reports the comparison. CycleSpectra achieves the best anatomical agreement (Dice 0.803, HD95 4.37 mm), the highest LV/RV curve correlation (0.727), the lowest SV error (32.81), and competitive LVEF error (16.60%, second to FusionNet's 12.40%, though FusionNet degrades heavily on anatomy with Dice 0.391 and curve correlation 0.159). UVI-Net retains the highest global PSNR/SSIM, which we attribute to its near-identity blending behavior preserving static background; under the heart-ROI and motion-fidelity columns, however, its motion is not well recovered (curve correlation 0.226). Dyna3DGR sits between the two regimes: per-pair test-time fitting makes it unaffected by the CT$\rightarrow$CT scanner shift (Dice 0.731, HD95 4.97 mm), but its motion-sensitive metrics again collapse (curve correlation 0.058, LVEF error 26.60%), consistent with its in-domain behavior reported in the main paper.

OOD-4DcMR (cross-modality). We further evaluate on the cMR cohort released by Yuan et al. [Yuan et al., 2023], which provides 24 cases at 25 phases each, 10 mm slice thickness. Chamber labels follow the multi-modality whole heart segmentation conventions established by [Zhuang et al., 2019]. This setting tests both modality shift (CT$\rightarrow$cMR) and the more aggressive thick-slab geometry. Voxel-level chamber masks are unavailable for this cohort, so we only report image fidelity and LV-mesh-driven functional metrics, and we restrict the comparison to the three methods that produce a deformation field applicable to the endpoint LV mesh: UVI-Net, Dyna3DGR, and CycleSpectra. CycleSpectra obtains the best nominal mean functional scores among the three methods: LV curve correlation is 0.743, compared with 0.729 for UVI-Net and 0.725 for Dyna3DGR; SV error is 64.16%, compared with 79.39% and 78.56%; and LVEF error is 15.60%, compared with 19.68% and 19.49%, respectively. The large across-subject variation in curve correlation (SD 0.407–0.492) indicates substantial cohort heterogeneity; we therefore interpret the curve-correlation ordering as nominal rather than evidence of statistical superiority. CycleSpectra trails UVI-Net by approximately 4.4 dB on global PSNR, again reflecting UVI-Net's near-identity behavior on static background regions. This cross-modality result is presented as a stress test rather than a primary comparison, given the substantial gap between the training distribution (CT, 1.6 mm thin slice) and the test distribution (cMR, 10 mm thick slabs).

Table A1: **Out-of-distribution two-frame inference.** All methods use the same fixed set of endpoint pairs and the same checkpoint trained only on the in-house 4D-CT cohort, with no retraining. *OOD-4DCT* uses 59 fixed opposite pairs from 4 external CT cases. *OOD-4DcMR* uses one fixed non-ED/ES opposite pair per case across 24 MR subjects; voxel-level chamber masks are unavailable, so we only report image fidelity and LV-mesh-driven functional metrics for the three methods that produce a deformation field applicable to the endpoint LV mesh. Bold indicates the best nominal mean and does not imply statistical significance.

Method	PSNR↑	SSIM↑	ROI PSNR↑	Dice avg↑	HD95↓	Curve corr↑	SV err.↓	LVEF err.↓
OOD-4DCT (4 cases; 59 fixed opposite pairs)								
FusionNet	17.55±0.99	0.597±0.073	15.81±1.11	0.391±0.081	20.88±3.82	0.159±0.131	39.54±4.55	**12.40**±2.00
TMKD	23.30±0.43	0.785±0.061	21.23±0.53	0.694±0.088	8.73±1.86	0.386±0.165	55.20±6.73	22.89±3.07
Mo-Diff	15.16±0.73	0.565±0.021	13.97±0.51	0.243±0.048	23.45±1.43	0.082±0.071	152.31±31.15	31.41±7.60
TTT4D	11.53±1.62	0.428±0.075	9.66±2.04	0.343±0.042	24.86±2.82	0.010±0.055	269.19±52.79	22.97±5.87
UVI-Net	25.87±1.07	0.912±0.026	22.95±0.87	0.627±0.148	10.76±2.40	0.226±0.118	50.32±5.44	25.03±2.69
Dyna3DGR	21.34±0.30	0.737±0.015	20.22±0.76	0.731±0.038	4.97±0.32	0.058±0.063	48.53±3.93	26.60±2.68
Ours	23.18±0.84	0.857±0.030	20.36±0.93	**0.803**±0.015	**4.37**±0.50	**0.727**±0.107	**32.81**±5.23	16.60±8.54
OOD-4DcMR (24 cases; one fixed non-ED/ES opposite pair per case)								
UVI-Net	33.75±2.00	0.973±0.006	—	—	—	0.729±0.492	79.39±12.68	19.68±5.95
Dyna3DGR	24.82±0.83	0.792±0.019	—	—	—	0.725±0.481	78.56±12.73	19.49±5.90
Ours	29.31±1.00	0.948±0.012	—	—	—	**0.743**±0.407	**64.16**±17.07	**15.60**±5.47

A.3 Robustness Analysis (Extended)

This section provides extended visualizations supporting the robustness claim discussed in the main paper. The main paper reports the LVEF-error heatmap over the opposite-pair phase grid (Fig. 3); here we report the corresponding heatmaps for two additional metrics, and the per-pair distributional view across the same 85-pair test bank.

Per-method heatmaps for ROI PSNR and Dice. Fig. A1 extends Fig. 3 with two additional panels: ROI PSNR (image fidelity within the heart ROI) and Dice average across chambers. Both metrics show the same qualitative pattern as LVEF error: methods relying on intensity interpolation (UVI-Net, TMKD) achieve favorable scores only near the canonical ED/ES intersection, while CycleSpectra remains comparatively uniform across most of the opposite-pair plane. The agreement across three independent metrics (image fidelity, anatomical overlap, and clinical-functional error) supports the conclusion that the spectral motion representation, rather than a pair-specific interpolation rule, is what drives performance away from the canonical extrema.

Per-pair distributional analysis. Fig. A2 shows violin plots of ROI PSNR, Dice, and |LVEF err.| across all 85 test pairs for each method. A narrow violin pinned at the favorable end of each axis corresponds to anchor-robust behavior. CycleSpectra shows a comparatively tight distribution on all three metrics (with the low-EF subject sub7 forming its upper LVEF tail), while UVI-Net and TMKD exhibit a long lower tail in Dice and a long upper tail in LVEF error despite competitive ROI PSNR — consistent with the heatmap finding that these methods perform well only on a subset of opposite-pair configurations.

A.4 Multi-Frame Extensions

This appendix presents two complementary multi-frame extensions. On the in-house 4D-CT cohort (Tab. A2), we apply post-hoc latent fusion to the frozen two-frame checkpoint without retraining. On the cross-modality 4D-cMR cohort (Tab. A3), we evaluate a separately trained adaptive variable-N extension of CycleSpectra trained on 4D-CT ($N\in\{2,3,4\}$) that mean-pools multi-path latent features before decoding. Neither setup uses cMR data for training or fine-tuning.

Inference-time latent fusion (Tab. A2). Given M input frames of one subject, we form all $M-1$ pairs sharing each frame as source and average their motion and context latents channel-wise; the query phase is decoded from the strict previous phase along the cycle. The two-frame checkpoint is kept frozen throughout.

Adaptive variable-N evaluation (Tab. A3). For each query phase, the phase-wise previous observed frame serves as source and remaining frames as contexts; the adaptive model processes the

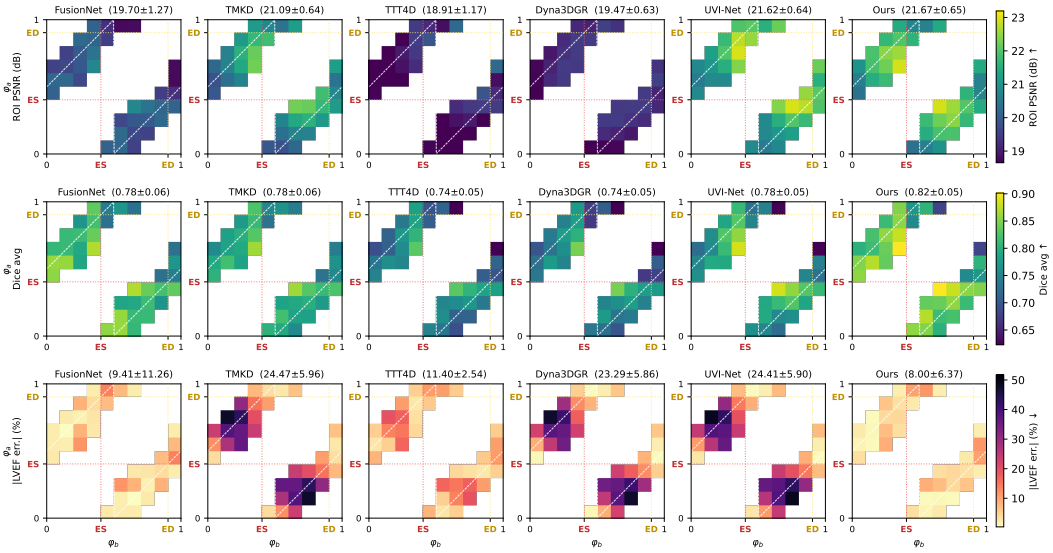

Figure A1: **Robustness to opposite-pair anchor choice (extended version of Fig. 3).** Per-method heatmaps over the (endpoint a, endpoint b) phase grid. Top: ROI PSNR. Middle: Dice average. Bottom: |LVEF err.| (same as Fig. 3). Dotted lines mark the cohort-mean canonical ED ($\bar{\psi}_{\mathrm{ED}} \approx 0.90$) and ES ($\bar{\psi}_{\mathrm{ES}} \approx 0.40$).

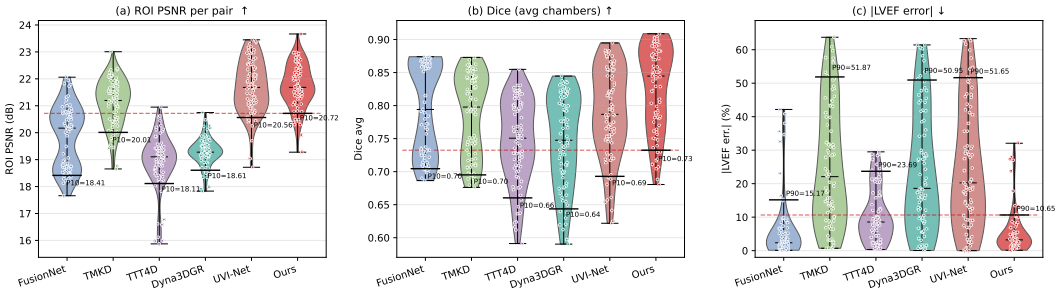

Figure A2: **Per-pair distribution of three robustness metrics.** Violin + scatter of all 85 opposite pairs per method. Black bars mark each method's worst-case percentile (P10 for higher-is-better metrics, P90 for |LVEF err.|). The dashed red line propagates our method's worst case across all columns.

shared source–context paths and mean-pools latent representations before decoding. For $M{=}2$, this reduces to one path, using the separately trained adaptive variable-N checkpoint rather than the two-frame checkpoint evaluated in Tab. A1.

Input-frame selection and evaluation protocol. For each subject, $M{=}2$ selects the opposite pair with cyclic distance nearest 0.5; $M{=}3$ and $M{=}4$ iteratively add the midpoint of the longer remaining arc. Image-fidelity metrics are computed on the common non-input frames across $M{\in}\{2,3,4\}$; full-cycle functional metrics fill observed input frames with ground truth. On 4D-cMR, voxel masks are unavailable, so image fidelity and LV-mesh functional metrics are reported. Because Tab. A1 evaluates the final two-frame checkpoint on held-out query frames while Tab. A3 evaluates the separate adaptive variable-N checkpoint under this common-frame protocol, their absolute $M{=}2$ values are not directly comparable; Tab. A3 is intended to measure the within-model scaling trend across M.

Results. On 4D-CT (Tab. A2), post-hoc latent fusion with the frozen two-frame model improves PSNR from 23.58 to 24.61 dB, Dice from 0.836 to 0.852, and LV/RV curve correlation from 0.846 to 0.941 as M increases from 2 to 4, confirming that the model supports test-time latent aggregation without retraining. On 4D-cMR (Tab. A3), the adaptive variable-N model exhibits a consistent

Table A2: **Multi-frame latent fusion: cohort summary on the in-house 4D-CT test set** (7 subjects, mean±std). Image-fidelity metrics are computed on common non-input frames across $M \in \{2, 3, 4\}$; cycle-level metrics are computed on the full cycle with input frames filled with ground truth.

M	PSNR↑	ROI PSNR↑	Dice avg↑	HD95↓	LV/RV curve corr↑	SV err. (%)↓	LVEF err. (%)↓
2	23.575±0.802	21.882±0.861	0.836±0.027	6.582±2.783	0.846±0.151	28.550±28.885	11.088±10.911
3	24.192±0.847	22.513±0.901	0.842±0.037	6.601±3.370	0.859±0.161	21.996±17.485	8.087±4.767
4	**24.606±1.077**	**22.921±1.004**	**0.852±0.044**	**5.859±2.320**	**0.941±0.034**	**15.546±12.196**	**6.347±4.188**

Table A3: **Adaptive variable-N CycleSpectra on the cross-modality 4DcMR cohort** (24 subjects, mean±std). Evaluates a separately trained adaptive multi-frame checkpoint rather than the two-frame checkpoint in Tab. A1. Image metrics are computed on common non-input frames across $M \in \{2, 3, 4\}$; functional metrics are computed on the full cycle with observed frames filled with ground truth. Comparisons are intended across M within this table, rather than between $M=2$ and Tab. A1.

M	PSNR↑	SSIM↑	NCC↑	MAE↓	RMSE↓	LV curve corr↑	LV SV err. (%)↓	LVEF err. (%)↓
2	22.232±0.637	0.872±0.012	0.875±0.019	0.0348±0.0037	0.0848±0.0062	-0.454±0.508	39.649±24.781	9.904±7.536
3	23.044±0.450	0.883±0.011	0.887±0.016	0.0326±0.0030	0.0808±0.0050	0.327±0.382	**17.484**±12.273	**3.510**±2.706
4	**24.011±0.530**	**0.896±0.010**	**0.906±0.012**	**0.0298±0.0029**	**0.0739±0.0042**	**0.534±0.214**	19.117±14.198	3.696±2.411

within-model scaling trend: PSNR increases from 22.23 to 24.01 dB, LV curve correlation improves from -0.454 to 0.534, and LVEF error drops from 9.90% to 3.70%. Both settings show that a few distributed observations effectively constrain the cyclic motion spectrum, with additional frames primarily refining image fidelity.

We frame these evaluations as extensions rather than main-paper claims; the two-frame setup in Sec. 4.3 remains the primary comparison point matching all baselines.

A.5 Baseline Adaptation and Mask Provenance

All methods receive the same two observed endpoint CT volumes at test time. Tab. A4 lists, for each method, its phase interface, training or fitting objective, test-time optimization, and the source of the predicted chamber mask. We do not claim identical architectures, phase representations, training objectives, or training-pair exposure.

Table A4: **Baseline adaptation and mask provenance.** Per-method phase handling, training objective, and source of the chamber mask used for anatomical and functional metrics.

Method	Phase interface	Training / fitting objective	Test-time opt.	Predicted-mask source
CycleSpectra	Analytic spectral read-out; anchor/context phases; bidirectional	Dense query-image reconstruction (ℓ_1, SSIM, foreground weighting), smoothness/folding, endpoint closure; no teacher	None	Warp source-endpoint label
UVI-Net	Arc-relative α; endpoints swapped on reverse arc	Bidirectional endpoint NCC + Charbonnier + flow regularization + native cycle-consistency; no GT intermediate frames	None	Hold source-endpoint label
FusionNet	Relative α; endpoints swapped on reverse arc	Direct reconstruction of sampled non-input query CT	None	Frozen nnU-Net re-segmentation
Mo-Diff	α-derived temporal differential field; whole query sequence jointly	Diffusion denoising of complete GT query sequence	None (200 sampling steps)	Frozen nnU-Net re-segmentation
Dyna3DGR	Deformation MLP queried at true phases	Per-pair fit, L_1 on the two endpoints only	**Yes:** 20,000 iters/pair, 12,288 nodes	Warp source-endpoint label
TMKD	Relative phase condition; forward/reverse heads	Distillation from dense motion teacher / direct frame reconstruction	None	Frozen nnU-Net re-segmentation
TTT4D	Test-time training token adaptation	Self-supervised test-time training	Adaptation on input frames	**Not verified**
Identity	Phase-wise endpoint-source selection	None	None	Hold selected source-endpoint label
T0 linear-flow	$F_{s\to q} = (V^{\mathrm{GT}}_{\mathrm{norm}}(s) - V^{\mathrm{GT}}_{\mathrm{norm}}(q))F_{\mathrm{ED}\to\mathrm{ES}}$	No learned training; precomputed ED–ES displacement	None	Warp source-endpoint label with $F_{s\to q}$

A.6 Local Motion Fidelity and Common-Segmenter Control

This section provides extended evaluation of local motion regularity and a controlled per-chamber comparison where a single frozen segmenter is applied uniformly to all methods.

Deformation regularity. Tab. A5 evaluates the Jacobian determinant $\det(I + \nabla u)$ for the two methods that provide a queryable continuous deformation field: CycleSpectra and Dyna3DGR. CycleSpectra attains near-incompressible and essentially fold-free deformation (0.009% folded voxels, mean $\det J$ 0.992, volume distortion 0.082), whereas Dyna3DGR exhibits a $2.4\times$ larger volume distortion (0.200). Note that UVI-Net provides no queryable displacement field, and FusionNet, TMKD, and Mo-Diff generate intensity volumes directly without standardized flow outputs.

Table A5: **Deformation regularity** (opposite-pair protocol; evaluated on methods providing a queryable displacement field). Volume distortion is defined as the spatial mean of $|\det(I+\nabla u)-1|$.

Method	Fold voxels (%) ↓	Mean $\det J$ (ideal 1)	Volume distortion ↓	Flow mag. (vox)
CycleSpectra	0.009	0.992	**0.082**	0.643
Dyna3DGR	**0.008**	1.010	0.200	2.515

Common-segmenter per-chamber evaluation. To control for differences in chamber mask provenance (Tab. A4), we re-score all seven methods on the identical 85 pairs / 7 subjects / 1,038 frames using a single frozen no-TTA nnU-Net [Isensee et al., 2021] applied to their generated CT volumes. Tab. A6 reports per-chamber Dice and Tab. A7 reports per-chamber HD95. CycleSpectra achieves the best cardiac-5 macro Dice (0.811) and HD95 (9.03 mm), leading on LV and myocardium, while isolated right-heart distance metrics are led by specific baselines.

Table A6: **Per-chamber Dice under a common frozen no-TTA nnU-Net segmenter** (opposite-pair protocol; all seven methods on the same 85 pairs / 7 subjects / 1,038 non-input frames; aggregated frame → pair → subject → cohort, mean ± population SD across subjects). Cardiac-5 is the unweighted macro-average over the five structures.

Method	LV	MYO	RV	LA	RA	Cardiac-5
CycleSpectra	**0.870**±0.031	**0.847**±0.032	**0.778**±0.087	**0.850**±0.040	0.708±0.124	**0.811**
UVI-Net	0.817±0.034	0.809±0.047	0.774±0.079	0.844±0.046	**0.712**±0.124	0.791
TTT4D	0.745±0.032	0.739±0.057	0.756±0.052	0.787±0.049	0.674±0.098	0.740
TMKD	0.804±0.037	0.792±0.052	0.766±0.080	0.838±0.046	0.706±0.118	0.781
FusionNet	0.849±0.050	0.813±0.028	0.747±0.077	0.841±0.047	0.679±0.128	0.786
Dyna3DGR	0.801±0.038	0.788±0.043	0.753±0.077	0.837±0.049	0.699±0.117	0.776
Mo-Diff†	0.058±0.014	0.059±0.017	0.064±0.012	0.060±0.018	0.060±0.014	0.060

†Mo-Diff complete cardiac-5 surface coverage is only 8.1%; segmentation failures are assigned Dice= 0.

Table A7: **Per-chamber HD95 (mm) under a common frozen no-TTA nnU-Net segmenter** (same protocol as Tab. A6). Lower is better.

Method	LV	MYO	RV	LA	RA	Cardiac-5
CycleSpectra	**5.29**±3.40	**5.09**±2.26	12.93±8.84	**7.77**±4.04	14.08±6.18	**9.03**
UVI-Net	6.67±2.77	5.56±1.82	12.61±8.27	7.86±4.05	14.08±6.11	9.36
TTT4D	10.26±5.12	7.59±2.53	**9.70**±3.21	9.05±4.00	**12.17**±5.26	9.75
TMKD	7.13±2.77	5.86±1.82	13.17±8.44	8.00±4.01	14.21±6.04	9.67
FusionNet	8.89±11.30	7.41±5.65	17.28±8.76	14.38±16.30	13.88±4.63	12.37
Dyna3DGR	7.06±2.63	5.44±1.02	14.33±7.44	18.65±25.56	27.59±29.84	14.61
Mo-Diff†	15.34±11.02	6.62±1.68	7.47±1.55	10.21±5.38	10.97±3.90	16.05†

†HD95 for Mo-Diff is computed only on valid surfaces; the other six methods have 100% valid surfaces. This control standardizes measurement across methods without altering each method's native input, architecture, or training objective.

A.7 Subject-Level Uncertainty and Statistical Reporting

Because the evaluation cohort comprises seven held-out subjects, we report subject-level 95% percentile bootstrap confidence intervals (10,000 resamples, random seed 20260727) for both evaluation protocols. Tab. A8 reports opposite-pair intervals, Tab. A9 reports canonical ED/ES intervals,

and Tab. A10 reports paired method-difference intervals (CycleSpectra − baseline) used to assess statistical superiority.

Table A8: **Opposite-pair protocol: subject-level 95% bootstrap confidence intervals** (mean [95% CI], 10,000 resamples). Metrics are aggregated frame → pair → subject.

Method	ROI PSNR	Dice	HD95 (mm)	Curve corr	SV err. (%)	LVEF err. (%)
CycleSpectra	21.67 [21.21, 22.14]	0.821 [0.784, 0.858]	7.94 [5.78, 10.05]	0.862 [0.806, 0.916]	44.12 [36.31, 53.49]	8.00 [4.15, 13.16]
UVI-Net	21.62 [21.08, 21.94]	0.785 [0.750, 0.819]	8.87 [6.89, 10.78]	0.102 [−0.058, 0.208]	57.39 [51.82, 61.98]	24.41 [19.49, 27.86]
TTT4D	18.91 [17.95, 19.63]	0.740 [0.705, 0.776]	10.29 [8.33, 12.28]	0.185 [0.055, 0.274]	46.83 [41.52, 52.54]	11.40 [9.61, 13.38]
TMKD	21.09 [20.56, 21.45]	0.778 [0.735, 0.820]	10.62 [7.97, 13.54]	0.187 [0.059, 0.306]	64.53 [59.15, 70.01]	24.47 [19.52, 27.97]
FusionNet	19.70 [18.80, 20.65]	0.779 [0.736, 0.825]	13.60 [8.71, 19.72]	0.853 [0.747, 0.925]	46.03 [25.89, 76.44]	9.41 [2.84, 18.60]
Mo-Diff	14.59 [13.82, 15.20]	0.061 [0.050, 0.070]	13.13 [8.21, 20.95]	0.196 [0.024, 0.346]	97.56 [70.04, 120.65]	43.00 [35.73, 52.08]
Dyna3DGR	19.47 [18.97, 19.90]	0.739 [0.701, 0.772]	9.83 [7.98, 11.60]	0.113 [−0.048, 0.219]	55.80 [50.27, 60.52]	23.29 [18.41, 26.61]

Table A9: **Canonical ED/ES protocol: subject-level 95% bootstrap confidence intervals** (mean [95% CI], 10,000 resamples).

Method	ROI PSNR	Dice	HD95 (mm)	Curve corr	SV err. (%)	LVEF err. (%)
CycleSpectra	21.09 [20.59, 21.70]	0.800 [0.754, 0.846]	9.83 [5.87, 15.10]	0.632 [0.446, 0.798]	38.03 [22.78, 57.99]	7.80 [3.39, 15.17]
UVI-Net	22.29 [21.63, 22.94]	0.751 [0.702, 0.800]	11.04 [7.27, 16.12]	−0.093 [−0.253, 0.095]	38.12 [20.79, 56.14]	7.32 [3.10, 11.82]
TTT4D	19.39 [18.44, 20.21]	0.719 [0.672, 0.765]	12.14 [8.39, 17.24]	−0.031 [−0.247, 0.171]	49.85 [35.71, 72.14]	5.52 [2.21, 9.15]
TMKD	21.00 [20.41, 21.58]	0.778 [0.734, 0.821]	10.54 [7.76, 13.65]	0.248 [0.088, 0.413]	42.15 [32.56, 51.59]	6.37 [3.00, 10.11]
FusionNet	19.71 [19.00, 20.45]	0.778 [0.734, 0.824]	13.32 [8.13, 20.00]	0.781 [0.635, 0.896]	41.44 [22.27, 74.04]	9.01 [2.27, 20.25]
Mo-Diff	14.24 [12.63, 15.64]	0.097 [0.057, 0.137]	7.46 [6.23, 9.32] (n=5)	0.143 [−0.062, 0.348] (n=5)	156.52 [120.40, 192.70]	43.00 [35.73, 52.08]
Dyna3DGR	19.66 [19.08, 20.24]	0.713 [0.666, 0.756]	11.96 [8.26, 16.89]	−0.060 [−0.228, 0.135]	37.57 [20.47, 55.34]	6.90 [3.58, 10.47]

Table A10: **Paired bootstrap differences (CycleSpectra − baseline) under opposite-pair evaluation** (10,000 subject resamples; * denotes 95% CI excludes zero).

Metric	vs. UVI-Net	vs. TTT4D	vs. TMKD	vs. FusionNet	vs. Mo-Diff	vs. Dyna3DGR
ROI PSNR	$+0.05$ $[-0.34, +0.41]$	$+2.77^*$ $[+2.20, +3.43]$	$+0.58^*$ $[+0.26, +0.87]$	$+1.97^*$ $[+0.92, +3.08]$	$+7.09^*$ $[+6.62, +7.58]$	$+2.21^*$ $[+1.85, +2.52]$
Dice	$+0.036^*$ $[+0.028, +0.044]$	$+0.081^*$ $[+0.071, +0.092]$	$+0.043^*$ $[+0.022, +0.062]$	$+0.042^*$ $[+0.025, +0.064]$	$+0.761^*$ $[+0.723, +0.801]$	$+0.082^*$ $[+0.069, +0.100]$
HD95 (mm)	-0.92^* $[-1.12, -0.71]$	-2.35^* $[-2.68, -2.06]$	-2.68^* $[-4.99, -0.37]$	-5.66^* $[-10.15, -2.35]$	-5.19 $[-14.78, +1.29]$	-1.89^* $[-2.55, -1.40]$
Curve corr	$+0.760^*$ $[+0.627, +0.938]$	$+0.676^*$ $[+0.564, +0.817]$	$+0.674^*$ $[+0.523, +0.833]$	**+0.008** $[-0.072, +0.100]$	$+0.665^*$ $[+0.540, +0.807]$	$+0.749^*$ $[+0.614, +0.927]$
SV err. (%)	-13.27^* $[-21.00, -4.53]$	-2.72 $[-7.64, +3.63]$	-20.41^* $[-29.14, -10.55]$	-1.91 $[-24.81, +12.90]$	-53.44^* $[-74.63, -28.99]$	-11.68^* $[-19.26, -2.75]$
LVEF err. (%)	-16.41^* $[-22.88, -6.35]$	-3.40 $[-8.20, +2.91]$	-16.48^* $[-22.89, -6.44]$	-1.41 $[-6.28, +2.67]$	-35.00^* $[-39.26, -30.85]$	-15.29^* $[-21.72, -5.30]$

* indicates that the 95% bootstrap interval of the paired difference excludes zero. In the canonical ED/ES setting, curve correlation vs. FusionNet is -0.150 $[-0.393, +0.066]$ (FusionNet nominally higher), SV error vs. Dyna3DGR is $+0.47$ $[-9.75, +9.55]$, and LVEF error intervals vs. all non-diffusion baselines include zero.

A.8 Ablation on Loss Design

Tab. A11 provides the complete quantitative results for the loss component ablation. Removing foreground weighting hurts most, while removing endpoint closure gives a consistent degradation across both image fidelity and motion correlation.

A.9 Sparse Training Supervision

To test how much the dense training-phase supervision matters, we retrain an otherwise identical model in a deliberately extreme regime: each training subject contributes only ED, ES, and a single mid-arc query near the midpoint of the ED→ES arc (66 of 318 training phases, three per subject); all remaining phases are withheld from the training data path, and there is no teacher initialization or supervision. Architecture, data splits, schedule, and the two-frame test protocol are unchanged, and testing uses the same 85 opposite pairs (Tab. A12).

Paired subject-level bootstrap differences (strict − Full, 10,000 resamples): PSNR -2.223 dB $[-2.467, -1.979]$; ROI PSNR -1.913 $[-2.185, -1.643]$; Dice -0.029 $[-0.036, -0.023]$; HD95

Table A11: **Ablation on loss design** (opposite-pair bank, 7-subject mean ± population SD). All rows use the final spectral-core architecture and are trained from scratch under the same opposite-pair training protocol.

Variant	PSNR↑	ROI PSNR↑	Dice avg↑	HD95↓	LV/RV curve corr↑	SV err. (%)↓	LVEF err. (%)↓
Full loss	23.134±0.683	21.675±0.646	0.821±0.050	7.941±2.908	0.862±0.075	44.12±11.90	8.00±6.37
w/o endpoint closure	23.014±0.674	21.547±0.625	0.818±0.051	8.020±2.887	0.845±0.076	44.69±11.88	8.07±7.41
w/o SSIM loss	23.067±0.670	21.642±0.637	0.821±0.050	7.929±2.922	0.869±0.072	42.99±12.61	7.25±7.01
w/o smooth/fold loss	23.131±0.675	21.683±0.644	0.822±0.050	7.892±2.915	0.868±0.071	43.35±11.66	7.63±6.73
w/o foreground weighting	23.066±0.702	21.512±0.642	0.818±0.052	8.044±2.901	0.824±0.076	47.33±8.48	10.23±4.02

Table A12: **Sparse training supervision** (opposite-pair bank, 7-subject mean ± population SD).

Setting	Visible phases	PSNR↑	ROI PSNR↑	Dice avg↑	HD95 (mm)↓	LV/RV curve corr↑	LVEF err. (%)↓
Full	318/318	23.134±0.683	21.675±0.646	0.821±0.050	7.941±2.907	0.862±0.075	7.995±6.368
ED+ES+one-mid (strict)	66/318	20.912±0.627	19.762±0.691	0.792±0.046	8.690±2.715	0.432±0.138	14.560±2.593

$+0.749$ mm [$+0.569$, $+0.932$]; curve correlation -0.430 [-0.523, -0.336]; LVEF error $+6.565$ pp [-0.494, $+11.318$]. The PSNR drop is substantial; the functional metrics degrade most, while the LVEF-error increase has an interval that includes zero.

A.10 Evaluation Audit Note and Extended Discussion

Evaluator standardization. We standardized aggregation to frame → pair → subject → cohort and fixed a duplicated LV-volume entry that affected only the endpoint-excluded Curve correlation and SV error (image fidelity, Dice, HD95 and LVEF are unaffected by construction); TTT4D is aggregated over subjects like all other methods. All in-domain benchmark, uncertainty, and ablation tables use the corrected, standardized evaluator. The single full-model opposite-pair curve correlation is 0.862.

Extended discussion of limitations. *Cohort size and verifiability:* The in-house cohort has 36 subjects (22/7/7); the 85 opposite pairs are nested in the same seven test subjects, so several comparisons, notably against FusionNet, are nominal-mean orderings with overlapping intervals. The raw CT cannot be shared under our data-use agreement. *Pathology and failure cases:* The cohort has no validated labels for arrhythmia, regional wall-motion abnormality, congenital disease, or post-operative anatomy, so we neither claim nor rule out robustness for these subgroups. The only low-ejection-fraction test subject (GT LVEF 32.3%, nearly flat volume–time curve) is our single clear failure (LVEF error 22.4% opposite-pair, 29.3% canonical; several baselines are lower on this subject, although our curve correlation remains the highest, 0.789): performance drops substantially on such abnormal-contraction hearts. *Dense training supervision:* "Sparse" refers to test-time conditioning only; training uses dense per-phase supervision, so the learned cohort-level motion prior depends on densely sampled training phases. *Stationary cycle:* We model one representative cycle as a stable periodic process, whereas real rhythms vary beat to beat and arrhythmic hearts depart from a fixed period; reading out the field against a real-time per-beat phase estimate is a direction for future work. *Evaluation:* Functional metrics rely on chamber labels whose provenance differs across baselines (Tab. A4; a common-segmenter control reproduces the ranking), and without ground truth for regional strain or displacement, local fidelity is assessed only through Jacobian regularity and per-chamber overlap. *Scope and uncertainty:* Patients with incomplete or abnormal cardiac function are under-represented; addressing this requires larger and more diverse (multi-center, pathological) cohorts. Finally, we model a deterministic reconstruction; modeling uncertainty over plausible trajectories is future work. As an untested hypothesis, the encoder's mapping from sparse 3D anatomy to per-site motion spectra may carry cues about motion function and could serve as a pretrained representation for coarse discrimination of cardiac function.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state three contributions: a spectral representation of cyclic cardiac motion, a two-frame phase-queryable reconstruction framework, and a whole-cycle motion benchmark on a thin-slice 4D cardiac CT cohort (Sec. 4.3–4.4, App. A.5–A.8). We state superiority only where paired subject-level intervals exclude zero, and limit claims to evaluated cohorts.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Sec. 5 discusses the small private cohort (36 subjects; seven test subjects with nested pairs), absence of validated pathology labels, one low-EF failure case, dense training-phase supervision (and the sparse-supervision control, Sec. 4.4 and App. A.9), the stationary-cycle assumption, label provenance and lack of regional-strain ground truth, and deterministic reconstruction.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper does not contain formal theorems requiring proof. The Fourier synthesis layer (Eq. (2)) is a closed-form analytic operation, with its assumptions (low-frequency cyclic motion) discussed in Sec. 3.2 and validated empirically in Sec. 4.4.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Architecture, training protocol, data sampling, losses (including endpoint closure, Sec. 3.4), and the canonical/opposite-pair protocols are specified in Sec. 3–4 and App. A.1; baseline adaptations and mask provenance are documented in App. A.5; evaluation aggregation and bootstrap procedure are in Sec. 4.2 and App. A.7.

   Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Code is available at `https://github.com/windrise/CycleSpectra`. The in-house clinical 4D-CT cohort cannot be publicly released under institutional data-use agreements.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Sec. 4.1–4.2 report the 22/7/7 split, optimizer and schedule, GPU configuration, data sampling, loss weights, and checkpoint selection (best validation PSNR; epoch 81); App. A.1 gives per-component widths and loss definitions, and App. A.9 the sparse-supervision control.

   Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: All main tables report the mean ± population SD over the seven test subjects (frame → pair → subject → cohort; pairs are not treated as independent). We additionally report 95% percentile bootstrap intervals over subjects (10,000 resamples) and paired bootstrap differences (App. A.7), and use paired intervals to decide where superiority is stated.

Guidelines:

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

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: Training uses four NVIDIA RTX 3090 GPUs with mixed precision and takes about 4 hours for 100 epochs; inference runs on a single RTX 3090; Dyna3DGR requires 20,000 per-pair fitting iterations.

Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: The research uses retrospectively collected, de-identified medical imaging data under institutional approval (Sec. 4.1) and no model is deployed in clinical decision-making.

Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Positive: whole-cycle motion reconstruction from sparse phases may eventually support functional assessment from fewer acquired phases. Negative: the model produces synthetic volumes that are not observed data, its accuracy degrades on abnormal-contraction hearts (Sec. 5), and it must not replace acquired phases in diagnostic decisions without clinical validation.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The released code has no plausible misuse pathway and does not contain patient data.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All assets are credited under their respective licenses: UVI-Net [Kim et al., 2024] (MIT, `github.com/phh0715/UVI-Net`); TransMorph [Chen et al., 2022] and VoxelMorph [Balakrishnan et al., 2019] (Apache-2.0); nnU-Net [Isensee et al., 2021] (Apache-2.0); Dyna3DGR [Fu et al., 2025] (MIT); 4D-cMR [Yuan et al., 2023] (CC BY-NC 4.0); and MM-WHS [Zhuang et al., 2019] (academic agreement).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The released code (at `https://github.com/windrise/CycleSpectra`) is accompanied by documentation and a license; private cohort data is not released.

    Guidelines:

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

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

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

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing or prospective research with human subjects.

Guidelines:

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

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

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

Answer: [Yes]

Justification: The retrospective use of de-identified 4D cardiac CT was approved by the institutional ethics committee of our partner hospital that provided the data, with informed consent waived due to the retrospective nature of the study.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs were used only for writing and editing assistance and do not affect the core methodology, scientific rigor, or originality of the research.

Guidelines:

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