FREE-BREATHING UNGATED CINE CARDIAC MRI IN SECONDS VIA A CROSS-SLICE TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

Cine cardiac MRI is the clinical gold standard for measuring heart function, but a full exam typically takes an hour or longer, since it depends on repeated breath-holds and electrocardiogram (ECG) gating that are difficult or unreliable for patients with dyspnea or arrhythmia. Real-time MRI captures each 2D frame within a fraction of a heartbeat, needing neither, but slices are acquired sequentially, each capturing the heart at a different respiratory position and cardiac state. Reassembling them into a coherent 3D volume, known as slice-to-volume reconstruction (SVR), still relies on breath-holds, ECG, or self-gating in existing methods, which also take minutes per subject. We present **CardioStitch**, to our knowledge the first cardiac SVR method that is simultaneously *free-breathing*, *gating- and synchronization-free*, and *ultra-fast*. Rather than synchronizing slices, we *condition* on a reference slice that defines the target cardiac state: a cross-slice transformer reasons jointly over the entire stack and, in a single forward pass, predicts dense 2D-to-3D deformations that correct respiratory and cardiac motion together, and varying the reference yields a 3D cine. Trained on 628 subjects from five public datasets, CardioStitch reconstructs a full cine in 2 s, over $35\times$ faster than optimization-based methods. It outperforms baselines given the full real-time series of every slice plus ECG or self-gating on all image metrics, cutting respiratory error from 4.2 to 0.9 mm, and on all functional metrics under simulated atrial fibrillation. CardioStitch also transfers without retraining to real free-breathing, ungated scans, including from a patient with atrial fibrillation.

1 INTRODUCTION

Cine cardiac magnetic resonance (CMR) is the reference standard for assessing cardiac function (Kramer et al., 2020). Ventricular volumes and ejection fraction are measured from a stack of 12–16 thick 2D short-axis slices, each assembled over many heartbeats using electrocardiography (ECG) gating while the patient holds their breath for every one or two slices (Christodoulou et al., 2024). Most other sequences in a clinical exam share these requirements, so a comprehensive exam typically takes 45–60 minutes or longer (Christodoulou et al., 2024; Raman et al., 2022), nearly half of it spent on breath-holds and recovery (Yang et al., 2025). The burden falls hardest on patients with heart failure, children, and those with arrhythmias such as atrial fibrillation (AF), for whom breath-holds are difficult and ECG gating is unreliable (Holtackers & Stuber, 2024).

Real-time imaging removes both constraints: each frame is acquired within a fraction of a heartbeat, during free breathing and without ECG (Zhang et al., 2010). However, slices are still acquired sequentially, each at a different respiratory position and cardiac state, so naively stacking them yields a distorted heart (Fig. 1). Recovering a coherent 3D heart from them is the problem of cardiac slice-to-volume reconstruction (SVR). Existing methods reintroduce the very constraints that real-time imaging removes (Table 1). General SVR methods (Uus et al., 2020; Ebner et al., 2020; Xu et al., 2023) assume a static anatomy and thus need gated, breath-hold input, while dynamic methods synchronize real-time frames across slices using ECG (Vogt et al., 2026) or self-gating (van Amerom et al., 2019), assuming that a given cardiac phase looks the same in every beat, which fails under arrhythmia. Most also optimize each subject for minutes, too slow to review and rescan while the patient is still in the scanner.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

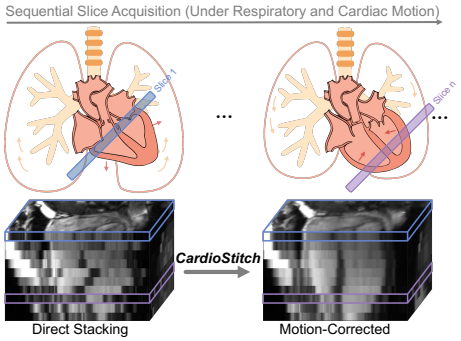

Figure 1: Motion-induced slice misalignment in free-breathing, ungated cardiac MRI, and our motion-corrected volume (right) via CardioStitch.

Method	Free breathing	No ECG / equipment	No self-gating/ synchronization	Runtime (s)↓
SVRTK	✗	✗	✗	167.4
NiftyMIC	✗	✗	✗	232.5
NeSVoR	✗	✗	✗	323.1
FC-SVR	✗	✗	✗	–
Dangi et al.	✗	✗	✗	**0.132**
Stolt-Ansó et al.	✗	✗	✗	–
Fetal CMR 4D†	✓	✓	✗	517.0
CiNeVol†	✓	✗	✗	70.10
Ours	✓	✓	✓	<u>1.962</u>

† Requires the full temporal series of every slice. Runtime is per full 3D cine.

Table 1: Comparison with SVRTK (Uus et al., 2020), NiftyMIC (Ebner et al., 2020), NeSVoR (Xu et al., 2023), FC-SVR (Young et al., 2024), Dangi et al. (2018), Stolt-Ansó et al. (2024), Fetal CMR 4D (van Amerom et al., 2019), and CiNeVol (Vogt et al., 2026).

We present **CardioStitch**, to our knowledge the first cardiac SVR method that is simultaneously (i) **free-breathing**, correcting respiratory motion instead of suppressing it; (ii) **gating- and synchronization-free**, needing no ECG or other equipment and never estimating or binning cardiac phases; and (iii) **ultra-fast**, reconstructing a full 3D cine in seconds. Instead of synchronizing slices, we *condition* on a reference slice that defines the target cardiac state. A VGGT-based transformer (Wang et al., 2025a) attends across the whole stack and, in a single forward pass, predicts dense 2D-to-3D deformations that jointly correct respiratory and cardiac motion; the aligned slices are splatted into a 3D volume, and varying the reference frame yields a 3D cine. Trained on 628 subjects from five public datasets, CardioStitch reconstructs a cine in about 2 s, more than 35× faster than optimization-based methods. It outperforms methods given strictly more input (the full real-time series of every slice, plus respiratory signals, ECG or self-gating) on all image metrics, reducing respiratory error from 4.2 to 0.9 mm, and on all functional metrics under simulated AF, and it transfers without retraining to real free-breathing, ungated scans of a healthy volunteer and an AF patient.

2 Related Work

Free-breathing and ungated cardiac MRI. Conventional cine CMR acquires each slice over multiple heartbeats with ECG gating and breath-holds (Kramer et al., 2020), which is difficult for patients with limited breath-hold capacity or irregular rhythms. Free-breathing methods relax respiratory control using navigators combined with ECG gating (Moghari et al., 2015; Pednekar et al., 2018), or remove external sensors by self-gating from the MR data (Larson et al., 2004; Feng et al., 2016), an idea extended to continuous free-running acquisitions (Di Sopra et al., 2019). However, these methods still estimate physiological states and retrospectively bin measurements, assuming that data assigned to the same state are consistent across heartbeats, which breaks down under arrhythmia. Real-time cine avoids binning altogether (Zhang et al., 2010), but each plane is reconstructed independently, so sequentially acquired slices remain at different respiratory and cardiac states.

Slice-to-volume reconstruction. SVR recovers a coherent volume from misaligned 2D slices, classically by iteratively registering multiple orthogonal stacks (Rousseau et al., 2006; Kuklisova-Murgasova et al., 2012; Ebner et al., 2020). SVRTK (Uus et al., 2020) adds deformable motion and NeSVoR (Xu et al., 2023) an implicit neural volume, but both optimize each subject. Learning-based methods predict slice poses (Hou et al., 2018; Xu et al., 2022) or, from a single stack, dense slice motion (Young et al., 2024). All of these assume a single static anatomy, which does not hold for the beating heart. In cardiac imaging, Dangi et al. (2018) correct only translational misalignment of gated cine stacks. Fetal CMR 4D (van Amerom et al., 2019) and CiNeVol (Vogt et al., 2026) instead acquire a real-time series at every slice and assign frames to cardiac states, using a constant per-slice heart rate or ECG and respiratory-belt signals, respectively; both are optimized per subject, and neither is designed for beat-to-beat variation. In contrast, CardioStitch reconstructs from free-breathing slices in a single forward pass, without gating or synchronization.

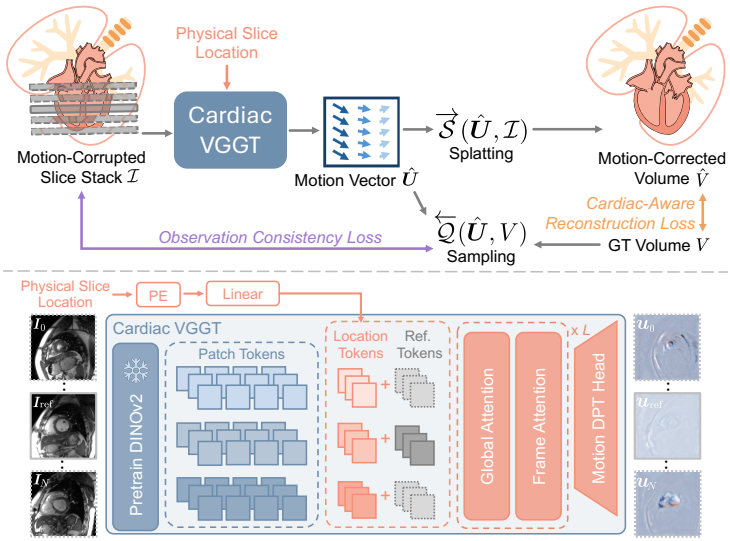

Figure 2: **Overview of CardioStitch (top).** Given motion corrupted slices and physical locations, Cardiac VGGT predicts dense 2D-to-3D motion fields, which are splatted into a motion corrected volume and trained with cardiac-aware reconstruction and observation consistency losses. **Structure of Cardiac VGGT (bottom)**, combining DINOv2 patch tokens with location and reference-state tokens, alternating frame and global attention, and a DPT head for dense 3D motion prediction.

3 Method

3.1 Problem Formulation

Let $\mathcal{I} = \{I_i\}_{i=1}^{N}$ denote a stack of N short-axis slices sequentially acquired from a time-varying 3D cardiac anatomy under free-breathing, ungated acquisition. For each slice I_i acquired at time t_i, the forward slice-acquisition process can be formulated as:

$$\boldsymbol{I}_i = \mathcal{A}_i(V_{t_i}) + \boldsymbol{n}_i, \tag{1}$$

where V_{t_i} denotes the underlying 3D anatomy at time t_i, $\mathcal{A}_i$ denotes the spatial sampling and the slice point-spread function, and $\boldsymbol{n}_i$ denotes measurement noise. Since each slice provides only a partial observation of the dynamic 3D anatomy, independently recovering V_{t_i} is severely under-constrained. To address this, we model each V_{t_i} with respect to a common target volume V and a slice-specific deformation field $\boldsymbol{u}_i$, then the acquisition model can be reformulated as:

$$\boldsymbol{I}_i = \mathcal{A}_i(V, \boldsymbol{u}_i) + \boldsymbol{n}_i. \tag{2}$$

The $\boldsymbol{u}_i$ specifies, for each pixel $\boldsymbol{p} = (x, y)$ in $\boldsymbol{I}_i$, the displacement vector $\boldsymbol{u}_i(\boldsymbol{p}) = (\Delta x, \Delta y, \Delta z)_i \in \mathbb{R}^3$ that maps it to the corresponding location in the target volume space. Cardiac SVR can therefore be formulated as estimating the set of dense 2D-to-3D correspondence fields $\boldsymbol{U} = \{\boldsymbol{u}_i\}_{i=1}^{N}$ from the asynchronous slices $\mathcal{I}$.

3.2 Cardiac VGGT

Conventional SVR methods (Uus et al., 2020; Xu et al., 2023; Ebner et al., 2020) typically recover slice-to-volume transformations through iterative, subject-specific optimization. In contrast, we directly predict dense displacement fields for all slices in a single forward pass with a feed-forward network. Specifically, building on VGGT (Wang et al., 2025a), a large feed-forward transformer designed for joint geometric reasoning across multiple views, we introduce **Cardiac VGGT**. Given an asynchronous slice stack $\mathcal{I}$, its physical slice locations $\mathcal{G}$ and a designated reference slice $\boldsymbol{I}_r$, Cardiac VGGT $\mathcal{F}_\theta$ directly predicts

$$\hat{\boldsymbol{U}} = \mathcal{F}_\theta(\mathcal{I}, \mathcal{G}, r), \tag{3}$$

where $\hat{\boldsymbol{U}} = \{\hat{\boldsymbol{u}}_i\}_{i=1}^{N}$ and r is the index of the reference slice.

As illustrated in Fig. 2 (bottom), each input slice $\boldsymbol{I}_i$ is first patchified into a set of K tokens $\boldsymbol{t}_i^I \in \mathbb{R}^{K \times C}$ through a frozen pretrained DINOv2 (Oquab et al., 2024). While these image tokens encode the observed cardiac appearance, they do not explicitly provide information about either the physical acquisition location or the target cardiac state toward which the stack should be aligned.

We therefore augment each slice representation with two additional tokens: a *physical-location token*, which encodes its physical acquisition location, and a *reference-state token*, which indicates whether the slice defines the target cardiac state. We describe their construction below.

Physical-location Token A discrete slice index captures only slice ordering and does not reflect the physical distance between slices, which may vary across acquisition protocols and datasets. We therefore condition each slice $\boldsymbol{I}_i$ on its physical through-plane location $z_i \in \mathcal{G}$. Specifically, z_i is first mapped using a Fourier embedding (Tancik et al., 2020) $\gamma(\cdot)$ followed by a learned linear projection to the token dimension $\boldsymbol{t}_i^g = \text{Linear}(\gamma(z_i)) \in \mathbb{R}^C$.

Reference-state Token While the physical-location token specifies where each slice is acquired, it does not specify the cardiac state to which the asynchronous stack should be aligned. We therefore designate a reference slice I_{ref}, whose observed anatomy defines the target cardiac state. Specifically, we choose the central slice of the stack as the reference, as it typically captures a larger portion of the heart and therefore provides a more informative representation of the cardiac configuration. To indicate the reference slice to the Cardiac VGGT, we introduce a reference-state token

$$\boldsymbol{t}_i^r = \begin{cases} \boldsymbol{e}_{\text{ref}}, & i = r, \\ \boldsymbol{e}_{\text{other}}, & i \neq r. \end{cases} \tag{4}$$

where $\boldsymbol{e}_{\text{ref}}, \boldsymbol{e}_{\text{other}} \in \mathbb{R}^C$ are learnable embeddings, and r denotes the index of the designated reference slice. The reference slice is assigned the dedicated embedding $\boldsymbol{e}_{\text{ref}}$, while all remaining slices share the same embedding $\boldsymbol{e}_{\text{other}}$.

Joint Reasoning and Motion Prediction For each slice, we combine the physical-location and reference-state tokens by element-wise addition, $\boldsymbol{t}_i^{\text{cond}} = \boldsymbol{t}_i^g + \boldsymbol{t}_i^r$, and concatenate the resulting slice-level token with the corresponding image patch tokens, $\widetilde{\boldsymbol{t}}_i = [\boldsymbol{t}_i^{\text{cond}}; \boldsymbol{t}_i^I]$. The conditioned tokens from all slices are jointly processed by the alternating frame-wise and global attention blocks of VGGT, enabling correspondence estimation for each slice to leverage information from the entire stack. Frame-wise attention captures intra-slice spatial context, while global attention enables information exchange across the entire stack, allowing the correspondence of each slice to be inferred in the context of all observations. Finally, a motion DPT head maps the resulting features back to the image resolution and predicts a dense 3D displacement field $\hat{\boldsymbol{u}}_i(\boldsymbol{p}) \in \mathbb{R}^3$ for every pixel $\boldsymbol{p} \in \mathbb{R}^2$.

3.3 Differentiable Slice-Volume Operators

After Cardiac VGGT predicts dense 2D-to-3D motion fields, we introduce two complementary differentiable operators: slice-to-volume splatting for reconstructing a target volume, and volume-to-slice sampling for directly evaluating the predicted correspondences in the observation space.

Let $\boldsymbol{g}_i(\boldsymbol{p}) = (\boldsymbol{p}, z_i) \in \mathbb{R}^3$ denote the physical coordinate of pixel $\boldsymbol{p}$ in slice $\boldsymbol{I}_i$ before motion correction. Given the predicted displacement $\hat{\boldsymbol{u}}_i$, its corresponding location in the target volume is

$$\widetilde{\boldsymbol{g}}_i(\boldsymbol{p}) = \boldsymbol{g}_i(\boldsymbol{p}) + \hat{\boldsymbol{u}}_i(\boldsymbol{p}). \tag{5}$$

Slice-to-Volume Splatting. We reconstruct the target volume by forward-splatting all observed slice pixels into a common 3D coordinate space:

$$\hat{V} = \overrightarrow{\mathcal{S}}(\hat{\boldsymbol{U}}, \mathcal{I}), \tag{6}$$

where $\overrightarrow{\mathcal{S}}$ denotes the differentiable slice-to-volume splatting operator. Specifically, for each observed pixel $\boldsymbol{p}$ in slice I_i, let $w(\boldsymbol{q}, \widetilde{\boldsymbol{g}}_i(\boldsymbol{p}))$ denote its trilinear contribution from the predicted 3D location $\widetilde{\boldsymbol{g}}_i(\boldsymbol{p})$ to a target voxel $\boldsymbol{q}$. The accumulated intensity and coverage at $\boldsymbol{q}$ are then

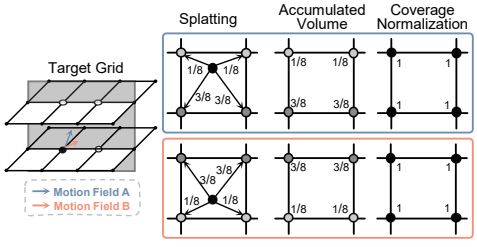

Figure 3: Illustration of normalization-induced ambiguity in forward splatting.

$$A(\boldsymbol{q}) = \sum_{i,\boldsymbol{p}} w(\boldsymbol{q}, \widetilde{\boldsymbol{g}}_i(\boldsymbol{p}))\, I_i(\boldsymbol{p}), \qquad C(\boldsymbol{q}) = \sum_{i,\boldsymbol{p}} w(\boldsymbol{q}, \widetilde{\boldsymbol{g}}_i(\boldsymbol{p}))\,. \tag{7}$$

The reconstructed voxel intensity is then obtained by coverage normalization,

$$\hat{V}(\boldsymbol{q}) = \frac{A(\boldsymbol{q})}{C(\boldsymbol{q}) + \epsilon}. \tag{8}$$

Although coverage normalization prevents the reconstructed intensity from being biased by nonuniform sampling density, it also introduces an ambiguity for correspondence estimation. As illustrated in Fig. 3, different predicted motions can produce different splatting weights and therefore different coverage patterns, while yielding the same normalized voxel intensities. For example, when neighboring observations have similar intensities, the motion-dependent interpolation weights appear in both the numerator and denominator of Eq. (8) and can largely cancel after normalization. This ambiguity is particularly pronounced for thick-slice acquisitions, where the through-plane observations are sparse and fewer independent measurements are available to constrain the correspondence.

Volume-to-Slice Sampling. To complement this aggregation-based constraint, we additionally introduce a volume-to-slice sampling operator. Given a target volume V, we obtain a set of resampled slices $\widetilde{\mathcal{I}} = \{\widetilde{I}_i\}_{i=1}^{N}$ as

$$\widetilde{\mathcal{I}} = \overleftarrow{\mathcal{Q}}(\hat{\boldsymbol{U}}, V), \tag{9}$$

where $\overleftarrow{\mathcal{Q}}$ denotes the differentiable volume-to-slice sampling operator. Specifically, each $\widetilde{I}_i(\boldsymbol{p})$ is obtained by sampling V at the predicted 3D location $\widetilde{\boldsymbol{g}}_i(\boldsymbol{p})$ using trilinear interpolation over the neighboring voxels.

Unlike splatting, sampling does not aggregate multiple observations and therefore does not involve coverage normalization. The sampled intensity depends directly on the predicted 3D coordinate, making the resulting observation-space constraint more sensitive to changes in correspondence wherever the target volume contains spatial intensity variation. Therefore, splatting and sampling provide complementary supervision: splatting enforces consistency of all observations in the reconstructed 3D space, whereas sampling constrains each predicted correspondence individually in the observation space. Pseudocode for both operators is given in Appendix A.7.

3.4 Training Objectives

Given the ground-truth target volume V in the end-expiratory respiratory state, we train CardioStitch with two complementary objectives. The *cardiac-aware reconstruction loss* supervises the splatted volume in the reconstruction space, encouraging coherent recovery of the target cardiac anatomy. The *observation consistency loss* instead samples V back to the observed slice space, providing direct supervision on the predicted 2D-to-3D correspondences.

Cardiac-Aware Reconstruction Loss. Because the heart occupies only a small fraction of the reconstructed volume, a whole-volume reconstruction loss can be dominated by surrounding anatomy that varies little across cardiac states, resulting in weak supervision for cardiac-state alignment. We therefore additionally emphasize the cardiac region using the paired cardiac mask M:

$$\mathcal{L}_{\mathrm{rec}} = \mathcal{L}_{\mathrm{vol}} + \lambda_h \mathcal{L}_{\mathrm{h}}, \qquad \mathcal{L}_{\mathrm{vol}} = \|\hat{V} - V\|_1, \qquad \mathcal{L}_{\mathrm{h}} = \|M \odot (\hat{V} - V)\|_1. \tag{10}$$

where M is the cardiac ROI mask associated with V, and λ_h controls the emphasis on the cardiac region. The cardiac-region term emphasizes cardiac-state-dependent structural variation, helping reduce the ambiguity between cardiac and respiratory motion during correspondence learning. Notably, the cardiac mask is required only during training and is not used at inference.

Observation Consistency Loss. As discussed in Sec. 3.3, coverage-normalized splatting can provide weak supervision for correspondence estimation under sparse through-plane sampling. This limitation is particularly relevant in the free-breathing setting, where through-plane motion is largely induced by respiration. We therefore directly enforce consistency in the observation space by sampling the ground-truth volume V at the predicted correspondences: $\widetilde{\mathcal{I}} = \overleftarrow{\mathcal{Q}}(\hat{\boldsymbol{U}}, V)$ and compare it with the acquired slices:

$$\mathcal{L}_{\mathrm{obs}} = \|\widetilde{\mathcal{I}} - \mathcal{I}\|_1. \tag{11}$$

Unlike splatting-based reconstruction supervision, this objective directly matches each observed pixel with the intensity sampled from the ground-truth volume at its predicted 3D location. This provides stronger supervision for through-plane localization, which is particularly important for correcting respiration-induced displacement.

We additionally apply an in-plane diffusion regularization (Balakrishnan et al., 2019) to the predicted displacement fields, encouraging spatially smooth and locally coherent correspondences within each

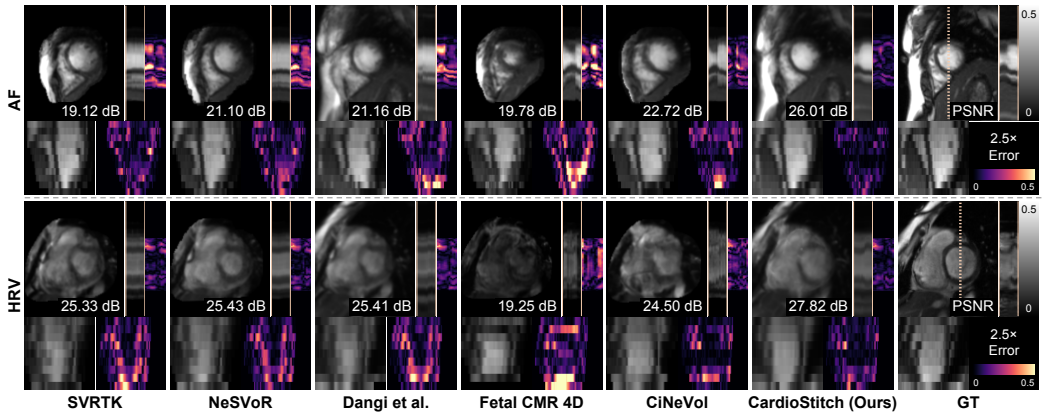

Figure 4: Qualitative comparisons for HRV and AF scenarios. For each method, we show the SAX reconstruction, its temporal profile along the dashed line, and the corresponding error map, together with the LAX reconstruction and its error map.

Table 2: Quantitative comparison of image reconstruction quality, respiratory motion accuracy, and runtime under AF and HRV. Best and second-best results are highlighted in **bold** and underlined.

Method	AF				HRV				Runtime (s)↓
	NCC↑	PSNR↑	SSIM↑	EPE_{resp}↓	NCC↑	PSNR↑	SSIM↑	EPE_{resp}↓	
SVRTK	0.783	21.82	0.578	4.791	0.787	21.94	0.584	4.801	167.4
NiftyMIC	0.790	20.03	0.558	4.695	0.794	20.09	0.562	4.709	232.5
NeSVoR	0.787	20.91	0.567	6.384	0.791	21.03	0.573	6.409	323.1
Dangi et al.	0.748	21.28	0.533	–	0.748	21.33	0.537	–	**0.132**
Fetal CMR 4D†	0.797	22.23	0.595	4.798	0.814	22.71	0.621	4.801	517.0
CiNeVol†	0.842	23.27	0.648	4.229	0.872	24.26	0.696	4.192	70.10
CardioStitch (Ours)	**0.883**	**24.50**	**0.715**	**0.917**	**0.884**	**24.59**	**0.719**	**0.916**	1.962

† Observes the full cine of every slice and additionally uses external gating (CiNeVol) or self-gating (Fetal CMR 4D).

slice. The regularization is imposed only along the two in-plane directions and does not enforce smoothness across sparsely sampled slice locations.

Finally, the overall training objective is $\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_o \mathcal{L}_{\text{obs}} + \lambda_s \mathcal{L}_{\text{s}}$, where $\mathcal{L}_{\text{s}}$ denotes the in-plane diffusion regularization, and λ_o and λ_s are balancing coefficients.

3.5 Inference

At inference, CardioStitch takes an asynchronous short-axis slice stack $\mathcal{I}$, its physical slice locations $\mathcal{G}$, and a designated reference slice $\boldsymbol{I}_r$ as input. In a single forward pass, Cardiac VGGT predicts the dense 2D-to-3D displacement fields $\hat{\boldsymbol{U}}$, which are subsequently mapped to the reconstructed target volume $\hat{V}$ using the forward-splatting operator $\overrightarrow{\mathcal{S}}$ defined in Eq. (6). The reconstructed volume follows the cardiac state specified by the reference frame while mapping the stack to the learned end-expiratory respiratory configuration.

To form a 3D cine, we keep the spatial reference fixed to the central (mid-ventricular) slice and vary only its temporal frame. Specifically, each frame of the central slice serves in turn as the reference, while one observed frame from every other slice is kept fixed across time. Each reference frame therefore specifies a target cardiac state and produces one reconstructed 3D volume, and the resulting sequence forms the 3D cine. No explicit cardiac or respiratory phase estimation, physiological gating, segmentation masks, or patient-specific optimization is required.

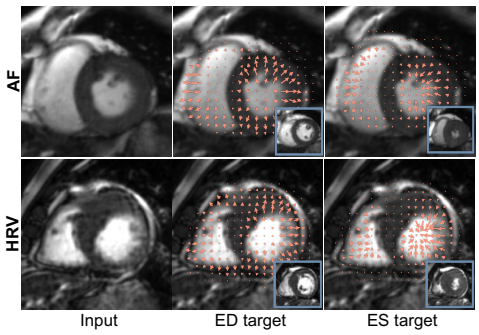

Figure 5: Visualization of learned cardiac motion under AF (top) and HRV (bottom). With all non-reference slices fixed, changing the reference from ED to ES changes the predicted deformation accordingly. Arrows show the in-plane cardiac motion after removing respiratory rigid shift.

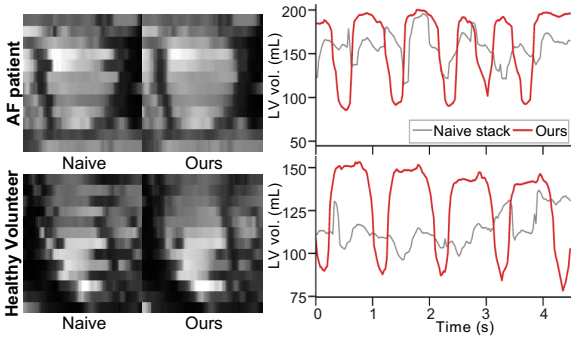

Figure 6: Prospective free-breathing, ungated reconstruction using CardioStitch in an AF patient (top) and a healthy volunteer (bottom). **Left:** LAX views from naive stacking and the CardioStitch motion-corrected volume. **Right:** LV volume over acquisition time for naive stacking (gray) and CardioStitch (red).

Table 3: Quantitative comparison of cardiac anatomical and functional accuracy under AF and HRV. Best and second-best results are highlighted in **bold** and underlined

Method	AF					HRV				
	Dice↑	HD_{95}↓	EF_{MAE}↓	LVM_{MAE}↓	VTC_{err}↓	Dice↑	HD_{95}↓	EF_{MAE}↓	LVM_{MAE}↓	VTC_{err}↓
SVRTK	0.703	11.26	35.69	14.44	20.29	0.706	11.04	38.45	15.89	18.19
NiftyMIC	0.701	11.57	39.05	15.66	20.57	0.704	11.28	40.81	15.42	18.00
NeSVoR	0.714	11.27	41.19	17.42	21.62	0.724	10.71	42.44	17.18	18.15
Dangi et al.	0.689	10.99	40.79	13.78	20.19	0.697	10.70	42.39	14.86	17.38
Fetal CMR 4D†	0.735	10.50	7.792	13.82	14.96	0.768	9.624	5.979	12.36	9.568
CiNeVol†	0.780	9.301	10.50	13.28	13.68	0.835	7.612	**5.582**	12.27	**6.541**
CardioStitch (Ours)	**0.847**	**6.954**	**7.490**	**9.108**	**7.258**	**0.847**	**6.883**	7.883	**9.471**	6.670

† Observes the full cine of every slice and additionally uses external gating (CiNeVol) or self-gating (Fetal CMR 4D).

4 Experimental Setup

Datasets. We combine five public ECG-gated breath-hold short-axis cine CMR datasets: CMRxRecon 2023/2024/2025 (Wang et al., 2024; 2025c;b), ACDC (Bernard et al., 2018), and M&Ms (Campello et al., 2021), spanning diverse acquisition and clinical settings (e.g., vendors, centres and cardiac pathologies). After quality control, 898 subjects are retained and split into 628/90/180 for training/validation/testing. Full details are provided in Appendix A.3.

Free-Breathing Ungated Simulation Process. To simulate free-breathing ungated scans, we model two major sources of inter-slice inconsistency: respiratory displacement and cardiac-state variation. We next describe 1) the respiratory motion model, 2) cardiac-state simulation, and 3) their use in the training and test protocols. **Respiratory motion**: We simulate a subject-specific 3D translational trajectory to the cine volume and extract each slice at an independently sampled respiratory phase while defining the reconstruction target at end expiration. Following the motion model of Lujan et al. (1999); Faranesh et al. (2013), the respiratory displacement is parameterized by phase r as $d_{\mathrm{SI}}(r) = A\sin^6(\pi r)$ and $d_{\mathrm{AP}}(r) = 0.35\, d_{\mathrm{SI}}(r)$, where $r = 0$ corresponds to end expiration and $A \sim \mathcal{U}(11.45, 26.15)$ mm specifies the subject-specific SI motion amplitude (Burger & Meintjes, 2013). The resulting displacement is further tilted by an angle sampled from $\mathcal{U}(0°, 45°)$ at a random azimuth to account for the oblique orientation of short-axis stacks relative to the body axes. **Cardiac-state simulation**: We use each subject's gated cine as the underlying cardiac motion trajectory rather than synthesizing additional cardiac deformation. Each cine is temporally represented by 12 cardiac phases spanning one cardiac cycle, with intermediate states obtained by linear interpolation between adjacent phases. These cardiac states are then sampled differently during training

and testing, as described below. **Training and test protocols**: *During training*, each slice is assigned an independently sampled cardiac state and respiratory phase to simulate an asynchronous free-breathing acquisition. The central slice serves as the reference, defining the target cardiac phase, with the corresponding end-expiratory, breathing-free volume as the reconstruction target. *At test time*, we simulate two cardiac-rhythm conditions: heart rate variability (HRV) to represent physiological beat-to-beat variation in normal subjects, and atrial fibrillation (AF) to represent pathological rhythm irregularity. Detailed rhythm simulation is described in Appendix A.4.

Baselines and metrics We compare our method with six representative methods: SVRTK (Uus et al., 2020), NiftyMIC (Ebner et al., 2020), NeSVoR (Xu et al., 2023), Dangi et al. (2018), Fetal CMR 4D (van Amerom et al., 2019) and CiNeVol (Vogt et al., 2026). Dangi et al. (2018) is retrained on our training split; baseline configurations are detailed in Appendix A.6.1. Notably, Fetal CMR 4D and CiNeVol take all temporal frames of every slice as input; Fetal CMR 4D self-gates the cardiac phase from these images, and CiNeVol is *additionally* provided with cardiac and respiratory states. We evaluate image quality, respiratory motion accuracy, anatomical and functional accuracy, and runtime. Full metric definitions and evaluation protocols are provided in Appendix A.6.

5 Results

5.1 Retrospective Results

Reconstruction quality and temporal consistency. As shown in Table 2, CardioStitch achieves the highest reconstruction accuracy under both AF and HRV. The temporal profiles in Fig. 4 further reveal that static SVR methods such as SVRTK and NeSVoR fail to resolve cardiac-state variation, while synchronization-based Fetal CMR 4D and CiNeVol show temporal discontinuities, especially under AF. In contrast, CardioStitch recovers temporally coherent cardiac motion without gating or synchronization.

Respiratory correction and learned cardiac motion. CardioStitch also substantially reduces the respiratory through-plane error, achieving an EPE of 0.917 mm under AF and 0.916 mm under HRV (Table 2). To visualize the learned cardiac motion, we perform a controlled experiment in Fig. 5: all non-reference slices are fixed, while only the reference frame is changed from ED to ES. After removing the slice-wise rigid respiratory shift, the predicted deformation shows corresponding ventricular expansion and contraction under both rhythms. Masking global attention at test time (Appendix A.2) shows that the model learns to correct the two motions separately: cardiac motion early in the network, by attending to the reference slice, and respiratory motion later, by attending among slices; masking either leaves the other intact.

Anatomical and functional accuracy. Table 3 shows that CardioStitch accurately preserves cardiac anatomy and function under AF, achieving the lowest errors across all reported metrics. Under HRV, where beat-to-beat variation is milder, Fetal CMR 4D and CiNeVol additionally benefit from dense temporal observations and additional motion information, resulting in lower EF errors; CiNeVol also attains a slightly lower VTC error. The LV volume–time curves in Fig. 7 further show that CardioStitch preserves the underlying cardiac dynamics without relying on cardiac synchronization.

Reconstruction efficiency. CardioStitch reconstructs a complete cardiac cine in seconds, substantially reducing reconstruction time compared with optimization-based methods. As illustrated in Fig. 7, it lies on the runtime–accuracy Pareto frontier under both rhythms, combining speed with accurate anatomical and functional reconstruction.

5.2 Evaluation on Prospective Real-Time Scans

All preceding experiments simulate free-breathing, ungated acquisition from gated breath-hold cine. We test CardioStitch, without retraining, on an in-house dataset comprising free-breathing, ungated real-time CMR from one healthy volunteer and one patient with atrial fibrillation (AF). Detailed acquisition parameters and ethical approval are provided in Appendix A.5. To form the cine, we follow Sec. 3.5 and vary only the reference frame of the central slice over time, while keeping all other slices fixed. As shown in Fig. 6, CardioStitch substantially improves both spatial and temporal consistency on prospective scans. In the LAX views (left), naive stacking exhibits pronounced

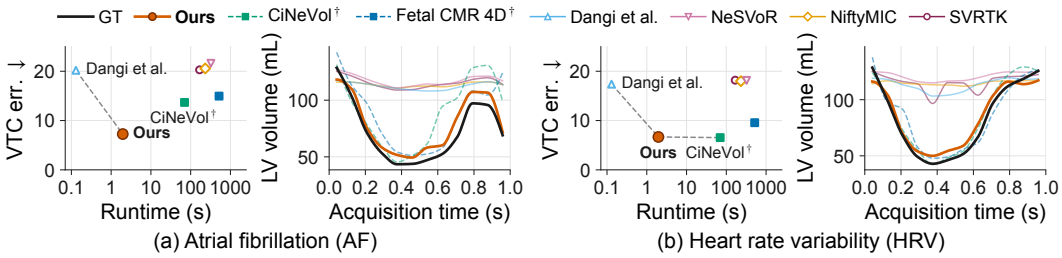

Figure 7: Runtime–accuracy trade-off on the test set and LV volume–time curves for one test subject under (a) atrial fibrillation and (b) heart rate variability. Within each panel, left: VTC error versus runtime, with the dashed Pareto frontier; right: reconstructed LV volume–time curves.

inter-slice discontinuities, whereas CardioStitch reconstructs a spatially coherent ventricle. The corresponding LV volume–time curves (right) show clear cardiac cycles after reconstruction, with relatively regular dynamics in the healthy volunteer and pronounced beat-to-beat variation in the AF patient. These dynamics are largely obscured by naive stacking.

5.3 Ablation Studies

We study the effect of each loss term in Table 4 and visualize the motion each model learns in Fig. A1. To quantify cardiac motion, we also report **Disp**$_{\text{heart}}$, defined as the mean in-plane displacement within the heart ROI after removing each slice's rigid respiratory shift.

Table 4: Ablation of the training objectives for AF.

$\mathcal{L}_h$	$\mathcal{L}_{\text{obs}}$	$\mathcal{L}_{\text{s}}$	PSNR↑	SSIM↑	VTC$_{\text{err}}$↓	Dice↑	EPE$_{\text{resp}}$↓	Disp.$_{\text{heart}}$
	✓		23.61	0.679	19.67	0.769	0.923	0.293
✓			23.73	0.680	17.66	0.783	2.417	1.411
✓	✓		**24.53**	0.711	7.553	0.845	**0.867**	1.599
✓	✓	✓	24.50	**0.715**	**7.258**	**0.847**	0.917	1.584

Observation consistency loss. Without $\mathcal{L}_{\text{obs}}$, the respiratory EPE rises from 0.917 to 2.417 mm, and the predicted through-plane shifts deviate from the true breathing (Fig. A1a). The heart still deforms (1.411 mm), but the VTC error rises to 17.66 and the LV volume curve flattens (Fig. A1c), so the recovered contraction is inaccurate. Supervising the correspondences directly in the observed slices is therefore needed for both respiratory and cardiac alignment.

Cardiac-aware reconstruction loss. Without the cardiac-region term (λ_h=0), breathing is still corrected (EPE 0.923 mm, Fig. A1a), but the in-plane deformation drops to 0.293 mm (Fig. A1b) and the heart barely contracts (Fig. A1c), raising the VTC error to 19.67. The model learns to correct breathing but not cardiac motion, consistent with the motivation in Sec. 3.4. **Smoothness regularization.** Adding $\mathcal{L}_{\text{s}}$ suppresses noisy local motion and further lowers the VTC error from 7.553 to 7.258, with small gains in Dice and SSIM. All trends hold under HRV (Appendix Table A1).

6 Conclusion & Limitations

We present **CardioStitch**, a feed-forward method that reconstructs a coherent 3D cardiac cine from free-breathing, ungated short-axis slices. Without ECG, self-gating, or per-subject optimization, it predicts dense 2D-to-3D correspondences for all slices in a single forward pass, correcting respiratory and cardiac motion together toward the cardiac state of a reference slice. It reconstructs a full 3D cine in seconds, remains accurate under atrial fibrillation, and transfers to real scans without retraining. Several limitations remain. CardioStitch uses a single short-axis stack, whose parallel slices leave the anatomy between slices underdetermined; long-axis views, routinely acquired in clinical exams, would constrain this through-plane geometry directly and could be added as further input slices. The reconstructed volume also keeps the slice spacing of the acquired stack, so its through-plane resolution remains limited by thick slices and their partial-volume effect. Reconstructing high-resolution isotropic volumes, by learning through-plane super-resolution jointly with motion correction, is a natural next step. Finally, our prospective evaluation covers two subjects, and a larger study across scanners and arrhythmias is needed to establish clinical reliability.

AI USE STATEMENT

In this work, we used generative AI tools (LLM-based coding assistants) to implement methods, including the training and evaluation code and the baseline reimplementations; to clean and reformat datasets; and to assist in analyzing and interpreting results. We have not used generative AI tools to propose or refine hypotheses, to develop the conceptual framework or formulation of our method, to design research methodology or experiments, or for translation, and tasks involving mathematical claims, proofs, or qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to provide feedback on the paper text. We have reviewed all AI-assisted work: code was tested and verified by the authors, and all reported results were re-checked against the underlying data. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The prospective real-time scans of one healthy volunteer and one patient with atrial fibrillation were acquired under an IRB-approved, HIPAA-compliant protocol with written informed consent from both participants (Appendix A.5). All other data are from publicly available, anonymized datasets.

REPRODUCIBILITY STATEMENT

All datasets used for training and quantitative evaluation are publicly available (Sec. 4); the curation procedure and the construction of the data splits are described in Appendix A.3. The free-breathing and arrhythmia simulation is specified in Sec. 4 and Appendix A.4, the model architecture and training objectives in Sec. 3, and the training settings, evaluation protocols, and baseline configurations in Appendix A.6 and Appendix A.6.1.

REFERENCES

Julius Åkesson, Jonathan Edlund, Katarina Steding-Ehrenborg, and Einar Heiberg. Retrospectively synchronized time-resolved ventricular cine images from 2D real-time exercise cardiac magnetic resonance imaging. *Clinical Physiology and Functional Imaging*, 45(5):e70027, 2025. doi: 10.1111/cpf.70027.

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

Olivier Bernard, Alain Lalande, Clément Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, 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. doi: 10.1109/TMI.2018.2837502.

Ian Burger and Ernesta M. Meintjes. Elliptical subject-specific model of respiratory motion for cardiac MRI. *Magnetic Resonance in Medicine*, 70(3):722–731, 2013. doi: 10.1002/mrm.24502.

Victor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021. doi: 10.1109/TMI.2021.3090082.

Anthony G Christodoulou, Gastao Cruz, Ayda Arami, Sebastian Weingärtner, Jessica Artico, Dana Peters, and Nicole Seiberlich. The future of cardiovascular magnetic resonance: all-in-one vs. real-time (part 1). *Journal of Cardiovascular Magnetic Resonance*, 26(1):100997, 2024.

C. S. Chung, M. Karamanoglu, and S. J. Kovács. Duration of diastole and its phases as a function of heart rate during supine bicycle exercise. *American Journal of Physiology-Heart and Circulatory Physiology*, 287(5):H2003–H2008, 2004. doi: 10.1152/ajpheart.00404.2004.

Shusil Dangi, Ziv Yaniv, and Cristian A. Linte. Cine cardiac MRI slice misalignment correction towards full 3D left ventricle segmentation. In *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10576, pp. 1057607. SPIE, 2018. doi: 10.1117/12.2294936.

Lorenzo Di Sopra, Davide Piccini, Simone Coppo, Matthias Stuber, and Jérôme Yerly. An automated approach to fully self-gated free-running cardiac and respiratory motion-resolved 5D whole-heart MRI. *Magnetic Resonance in Medicine*, 82(6):2118–2132, 2019. doi: https://doi.org/10.1002/mrm.27898. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.27898.

Michael Ebner, Guotai Wang, Wenqi Li, Michael Aertsen, Premal A. Patel, Rosalind Aughwane, Andrew Melbourne, Tom Doel, Steven Dymarkowski, Paolo De Coppi, Anna L. David, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain MRI. *NeuroImage*, 206:116324, 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.116324. URL https://www.sciencedirect.com/science/article/pii/S1053811919309152.

Anthony Z. Faranesh, Peter Kellman, Kanishka Ratnayaka, and Robert J. Lederman. Integration of cardiac and respiratory motion into MRI roadmaps fused with X-ray. *Medical Physics*, 40(3):032302, 2013. doi: 10.1118/1.4789919.

Li Feng, Leon Axel, Hersh Chandarana, Kai Tobias Block, Daniel K. Sodickson, and Ricardo Otazo. XD-GRASP: Golden-angle radial MRI with reconstruction of extra motion-state dimensions using compressed sensing. *Magnetic Resonance in Medicine*, 75(2):775–788, 2016. doi: 10.1002/mrm.25665.

Peter M. Full, Fabian Isensee, Paul F. Jäger, and Klaus Maier-Hein. Studying robustness of semantic segmentation under domain shift in cardiac MRI. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, volume 12592 of *Lecture Notes in Computer Science*, pp. 238–249, Cham, 2021. Springer. doi: 10.1007/978-3-030-68107-4_24.

Florian Fürnrohr and Reinhard Heckel. Piecewise dynamic diffusion regularization for reconstruction of cardiac cine MRI. *arXiv preprint arXiv:2607.03299*, 2026. doi: 10.48550/arXiv.2607.03299.

A. T. M. Gosselink, P. K. Blanksma, H. J. G. M. Crijns, I. C. Van Gelder, P.-J. de Kam, H. L. Hillege, M. G. Niemeijer, K. I. Lie, and F. L. Meijler. Left ventricular beat-to-beat performance in atrial fibrillation: contribution of Frank-Starling mechanism after short rather than long RR intervals. *Journal of the American College of Cardiology*, 26(6):1516–1521, 1995. doi: 10.1016/0735-1097(95)00340-1.

Robert J. Holtackers and Matthias Stuber. Free-running cardiac and respiratory motion-resolved imaging: A paradigm shift for managing motion in cardiac MRI? *Diagnostics*, 14(17), 2024. ISSN 2075-4418. doi: 10.3390/diagnostics14171946. URL https://www.mdpi.com/2075-4418/14/17/1946.

Benjamin Hou, Bishesh Khanal, Amir Alansary, Steven McDonagh, Alice Davidson, Mary Rutherford, Jo V. Hajnal, Daniel Rueckert, Ben Glocker, and Bernhard Kainz. 3-D reconstruction in canonical co-ordinate space from arbitrarily oriented 2-D images. *IEEE Transactions on Medical Imaging*, 37(8):1737–1750, 2018. doi: 10.1109/TMI.2018.2798801.

Christopher M. Kramer, Jörg Barkhausen, Chiara Bucciarelli-Ducci, Scott D. Flamm, Raymond J. Kim, and Eike Nagel. Standardized cardiovascular magnetic resonance imaging (CMR) protocols: 2020 update. *Journal of Cardiovascular Magnetic Resonance*, 22(1):17, 2020. ISSN 1097-6647. doi: https://doi.org/10.1186/s12968-020-00607-1. URL https://www.sciencedirect.com/science/article/pii/S1097664723002697.

Maria Kuklisova-Murgasova, Gerardine Quaghebeur, Mary A. Rutherford, Joseph V. Hajnal, and Julia A. Schnabel. Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. *Medical Image Analysis*, 16(8):1550–1564, 2012. doi: 10.1016/j.media.2012.07.004.

Andrew C. Larson, Richard D. White, Gerhard Laub, Elliot R. McVeigh, Debiao Li, and Orlando P. Simonetti. Self-gated cardiac cine MRI. *Magnetic Resonance in Medicine*, 51(1):93–102, 2004. doi: 10.1002/mrm.10664.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Anthony E. Lujan, Edward W. Larsen, James M. Balter, and Randall K. Ten Haken. A method for incorporating organ motion due to breathing into 3D dose calculations. *Medical Physics*, 26(5): 715–720, 1999. doi: 10.1118/1.598577.

Marek Malik, J. Thomas Bigger, A. John Camm, Robert E. Kleiger, Alberto Malliani, Arthur J. Moss, and Peter J. Schwartz. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 17(3):354–381, 1996. doi: 10.1093/oxfordjournals.eurheartj.a014868.

M. Markl, D. C. Lee, J. Ng, J. C. Carr, and J. J. Goldberger. Impact of cardiac arrhythmia on velocity quantification by ECG gated phase contrast MRI. *Journal of Cardiovascular Magnetic Resonance*, 17(Suppl 1):Q16, 2015. doi: 10.1186/1532-429X-17-S1-Q16. Abstract.

Mehdi H. Moghari, Rukmini Komarlu, David Annese, Tal Geva, and Andrew J. Powell. Free-breathing steady-state free precession cine cardiac magnetic resonance with respiratory navigator gating. *Magnetic Resonance in Medicine*, 73(4):1555–1561, 2015. doi: https://doi.org/10.1002/mrm.25275. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.25275.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Amol S Pednekar, Hui Wang, Scott Flamm, Benjamin Y Cheong, and Raja Muthupillai. Two-center clinical validation and quantitative assessment of respiratory triggered retrospectively cardiac gated balanced-ssfp cine cardiovascular magnetic resonance imaging in adults. *Journal of Cardiovascular Magnetic Resonance*, 20(1):44, 2018.

Subha V Raman, Michael Markl, Amit R Patel, Jennifer Bryant, Bradley D Allen, Sven Plein, and Nicole Seiberlich. 30-minute CMR for common clinical indications: a society for cardiovascular magnetic resonance white paper. *Journal of Cardiovascular Magnetic Resonance*, 24(1):13, 2022.

Francois Rousseau, Orit A. Glenn, Bistra Iordanova, Claudia Rodriguez-Carranza, Daniel B. Vigneron, James A. Barkovich, and Colin Studholme. Registration-based approach for reconstruction of high-resolution in utero fetal MR brain images. *Academic Radiology*, 13(9):1072–1081, 2006. doi: 10.1016/j.acra.2006.05.003.

Nil Stolt-Ansó, Vasiliki Sideri-Lampretsa, Maik Dannecker, and Daniel Rueckert. Intensity-based 3D motion correction for cardiac MR images. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4. IEEE, 2024. doi: 10.1109/ISBI56570.2024.10635803.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020.

Alena Uus, Tong Zhang, Laurence H. Jackson, Thomas A. Roberts, Mary A. Rutherford, Joseph V. Hajnal, and Maria Deprez. Deformable slice-to-volume registration for motion correction of fetal body and placenta MRI. *IEEE Transactions on Medical Imaging*, 39(9):2750–2759, 2020. doi: 10.1109/TMI.2020.2974844.

Joshua F. P. van Amerom, David F. A. Lloyd, Maria Deprez, Anthony N. Price, Shaihan J. Malik, Kuberan Pushparajah, Milou P. M. van Poppel, Mary A. Rutherford, Reza Razavi, and Joseph V. Hajnal. Fetal whole-heart 4D imaging using motion-corrected multi-planar real-time MRI. *Magnetic Resonance in Medicine*, 82(3):1055–1072, 2019. doi: 10.1002/mrm.27798.

Nora Vogt, Karyna Isaieva, Jean-Sébastien Louis, Pierre-André Vuissoz, Jacques Felblinger, and Julien Oster. Cardiac neural representations for ECG-guided slice-to-volume reconstruction. *Medical & Biological Engineering & Computing*, 64(4):1505–1516, 2026. doi: 10.1007/s11517-025-03488-7.

Chengyan Wang, Jun Lyu, Shuo Wang, Chen Qin, Kunyuan Guo, Xinyu Zhang, Xiaotong Yu, et al. CMRxRecon: A publicly available k-space dataset and benchmark to advance deep learning for cardiac MRI. *Scientific Data*, 11(1):687, 2024. doi: 10.1038/s41597-024-03525-4.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a.

Zi Wang, Mingkai Huang, Zhang Shi, Hongjie Hu, Lan Lan, Hui Zhang, Yan Li, et al. Enabling ultra-fast cardiovascular imaging across heterogeneous clinical environments with A generalist foundation model and multimodal database. *arXiv preprint arXiv:2512.21652*, 2025b. doi: 10.48550/arXiv.2512.21652.

Zi Wang, Fanwen Wang, Chen Qin, Jun Lyu, Cheng Ouyang, Shuo Wang, Yan Li, et al. CMRxRecon2024: A multimodality, multiview k-space dataset boosting universal machine learning for accelerated cardiac MRI. *Radiology: Artificial Intelligence*, 7(2):e240443, 2025c. doi: 10.1148/ryai.240443.

Junshen Xu, Daniel Moyer, P. Ellen Grant, Polina Golland, Juan Eugenio Iglesias, and Elfar Adalsteinsson. SVoRT: Iterative transformer for slice-to-volume registration in fetal brain MRI. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 13436 of *Lecture Notes in Computer Science*, pp. 3–13. Springer, 2022. doi: 10.1007/978-3-031-16446-0_1.

Junshen Xu, Daniel Moyer, Borjan Gagoski, Juan Eugenio Iglesias, P. Ellen Grant, Polina Golland, and Elfar Adalsteinsson. NeSVoR: Implicit neural representation for slice-to-volume reconstruction in MRI. *IEEE Transactions on Medical Imaging*, 42(6):1707–1719, 2023. doi: 10.1109/TMI.2023.3236216.

Kai Yang, Chen Cui, Fei Teng, Gang Yin, Jing An, Xinling Yang, Jinghui Li, Xiaoming Bi, Jianing Pang, Kelvin Chow, et al. Full free-breathing cardiac MRI: Enhancing efficiency and image quality in clinical practice. *Journal of Cardiovascular Magnetic Resonance*, pp. 101955, 2025.

Sean I. Young, Yaël Balbastre, Bruce Fischl, Polina Golland, and Juan Eugenio Iglesias. Fully convolutional slice-to-volume reconstruction for single-stack MRI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11535–11545, 2024. doi: 10.1109/CVPR52733.2024.01096.

Shuo Zhang, Martin Uecker, Dirk Voit, Klaus-Dietmar Merboldt, and Jens Frahm. Real-time cardiovascular magnetic resonance at high temporal resolution: radial flash with nonlinear inverse reconstruction. *Journal of Cardiovascular Magnetic Resonance*, 12(1):39, 2010. ISSN 1097-6647. doi: https://doi.org/10.1186/1532-429X-12-39. URL `https://www.sciencedirect.com/science/article/pii/S1097664723013194`.

A Appendix

A.1 Ablation Analysis

Qualitative analysis of the learned motion. Figure A1 provides a representative example illustrating how the different objectives affect the motion estimated by CardioStitch. Without the cardiac-aware term $\mathcal{L}_{\mathrm{h}}$, the model accurately recovers the slice-wise respiratory shift, but predicts almost no in-plane deformation within the heart. Consequently, the reconstructed LV volume varies only weakly over time and substantially underestimates the underlying contraction. This indicates that whole-volume supervision alone is dominated by relatively static surrounding anatomy and provides insufficient signal for learning cardiac-state alignment. Removing the observation consistency loss $\mathcal{L}_{\mathrm{obs}}$ leads to the opposite behavior. The model predicts appreciable cardiac deformation, but the estimated through-plane respiratory shifts deviate substantially from the ground truth. The resulting LV volume curve captures some contraction but remains markedly attenuated. This supports the role of $\mathcal{L}_{\mathrm{obs}}$ in directly constraining slice-to-volume correspondences, particularly along the sparsely sampled through-plane direction. When both $\mathcal{L}_h$ and $\mathcal{L}_{\mathrm{obs}}$ are used, the model recovers both respiratory alignment and substantial cardiac deformation, producing an LV volume curve close to the ground truth. Adding the smoothness regularizer $\mathcal{L}_s$ further suppresses locally irregular displacement vectors while preserving the overall motion amplitude and cardiac dynamics. Together, these results show that the three objectives play complementary roles: $\mathcal{L}_{\mathrm{obs}}$ anchors respiratory localization, $\mathcal{L}_h$ promotes cardiac-state alignment, and $\mathcal{L}_s$ regularizes the spatial structure of the predicted motion.

Consistency under HRV. Table A1 shows that the same trends hold under heart-rate variability. Removing $\mathcal{L}_{\mathrm{obs}}$ increases the respiratory EPE from 0.916 to 2.404 mm, whereas removing $\mathcal{L}_h$ reduces the mean cardiac displacement from 1.506 to 0.293 mm and increases the VTC error from 6.670 to 16.78. Using both terms substantially improves image, anatomical, and functional accuracy, while $\mathcal{L}_s$ provides a further improvement in SSIM, Dice, and VTC error. The agreement between AF and HRV indicates that the roles of the individual objectives are not specific to a particular rhythm simulation.

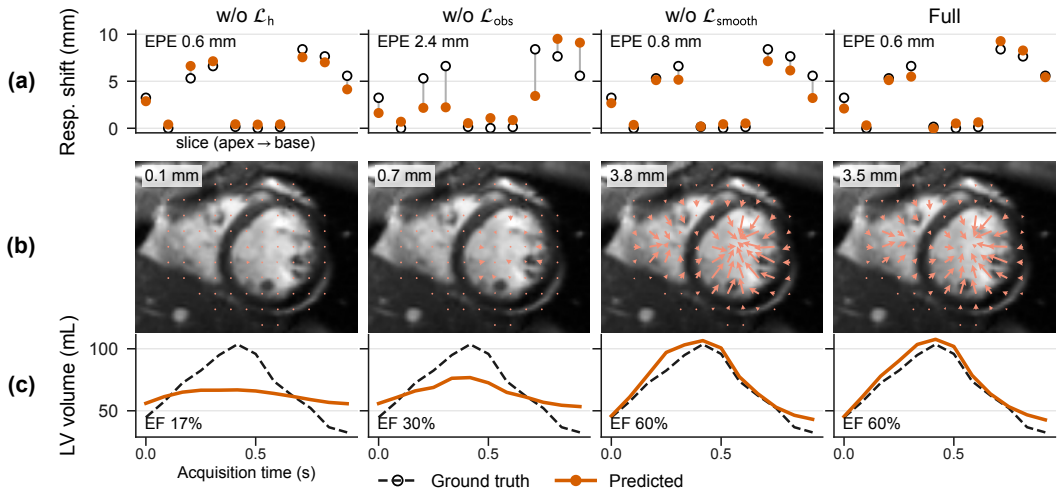

Figure A1: Motion learned by each ablated model on a representative test subject. (a) Predicted and ground-truth through-plane respiratory shift per slice. (b) In-plane deformation of one input slice, shown as arrows. The top-left number is the mean in-plane deformation magnitude within the heart ROI of this slice. (c) LV volume–time curve.

A.2 Attention Analysis

CardioStitch corrects cardiac and respiratory motion in a single forward pass. To see how the network handles the two, we examine its global attention. This is the only place where slices exchange

Table A1: Ablation of the training objectives for HRV.

$\mathcal{L}_h$	$\mathcal{L}_{\text{obs}}$	$\mathcal{L}_{\text{s}}$	PSNR↑	SSIM↑	VTC_{err}↓	Dice↑	EPE_{resp}↓	$\text{Disp.}_{\text{heart}}$
	✓		23.76	0.686	16.78	0.779	0.917	0.293
✓			23.87	0.686	15.21	0.792	2.404	1.410
✓	✓		**24.61**	0.715	6.970	0.846	**0.872**	1.518
✓	✓	✓	24.59	**0.719**	**6.670**	**0.847**	0.916	1.506

information. In each of the 24 global-attention blocks (numbered 0–23), every token attends to the tokens of all slices. At test time, without retraining, we mask the attention in selected blocks. The mask excludes selected slices from each token's softmax, so the token cannot read from them. We then observe which correction is lost.

We locate the relevant blocks by scanning. Masking the attention from the non-reference slices to the reference slice stops them from following the reference's cardiac state only in block 3. Masking any other single block leaves them unchanged. Masking all attention between slices in blocks 12–17 removes the respiratory correction completely. Blocks 0–11 remove part of it, and blocks 18–23 have no effect. We therefore compare two ablations. In *w/o reference attention (block 3)*, the non-reference slices cannot attend to the reference slice in block 3. In *w/o global attention (blocks 12–17)*, each slice attends only to itself in these blocks.

Where the reference is read. Fig. A2 shows the attention of one point on a non-reference slice. In block 3, it attends mainly to the heart of the reference slice. In other blocks, such as block 8, its attention is spread over the whole stack. In 94% of the 180 test subjects, block 3 is the block in which the non-reference slices attend most to the reference.

Heartbeat. Masking reference attention in block 3 alone suppresses the heartbeat of the non-reference slices. In Fig. A3(b), their LV volume swing drops from 75 to 12 mL (ground truth 111 mL), and over the test set their VTC error rises from 7.4 to 17.5 and their EF error from 8.8 to 48.7 pp, close to naive stacking (Table A2). Respiratory correction is unaffected. The same holds on the prospective scans of Sec. 5.2: the LV volume swing of the non-reference slices falls from 91.8 to 2.4 mL in the AF patient and from 52.0 to 1.7 mL in the healthy volunteer.

Breathing. Masking global attention in blocks 12–17 has the opposite effect. The predicted through-plane shifts drop to nearly zero (Fig. A3(c), bottom) and the respiratory error rises to that of naive stacking, while the heartbeat is kept (Fig. A3(c), top; Table A2). Respiratory motion is thus corrected by comparing slices with each other, not by reading the reference.

Together, CardioStitch reads the target cardiac state from the reference slice early in the network and corrects respiratory motion later by comparing slices, as intended by the reference-conditioned design.

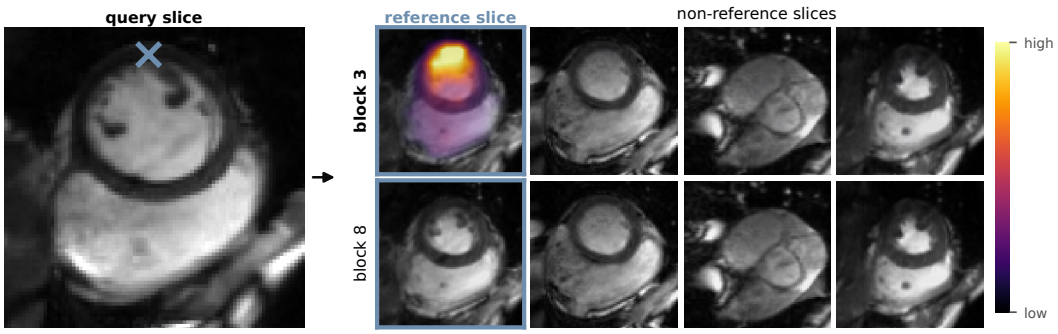

Figure A2: Global attention from one point (×) on a non-reference slice to the reference slice (framed) and three other slices, in global-attention blocks 3 (top) and 8 (bottom), averaged over the 16 heads (one test subject).

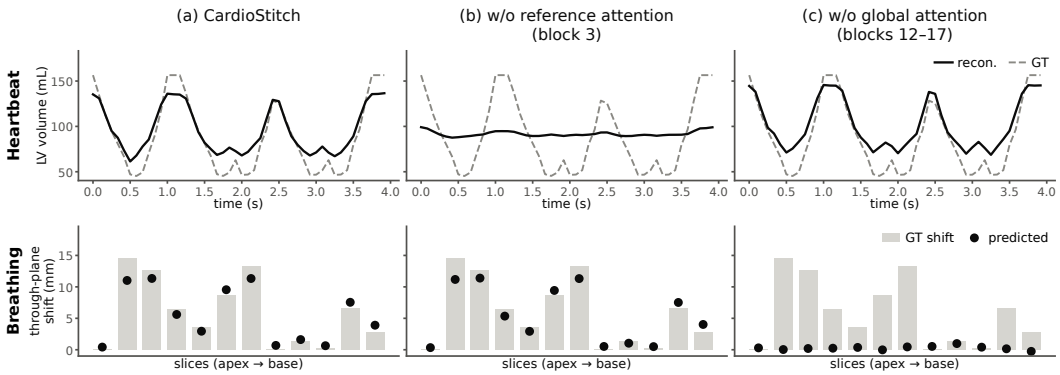

Figure A3: Test-time attention masking on one test subject under simulated AF. Top: LV volume over time, reconstructed (solid) and ground truth (dashed), summed over all slices except the reference slice. Bottom: simulated through-plane respiratory shift (bars) and predicted shift (dots) of each slice, in the first frame. Masking reference attention flattens the LV volume curve but keeps the respiratory correction (b); masking global attention in blocks 12–17 does the reverse (c).

Table A2: Test-time attention masking on the test set with gated cardiac timing. $\mathrm{VTC}_{\mathrm{err}}$ and $\mathrm{EF}_{\mathrm{MAE}}$ are computed on the non-reference slices. Naive stacking places the slices at their acquired positions without correction.

	$\mathbf{VTC}_{\mathrm{err}}\downarrow$	$\mathbf{EF}_{\mathrm{MAE}}\downarrow$	$\mathbf{EPE}_{\mathrm{resp}}\downarrow$
CardioStitch	7.41	8.82	0.98
w/o reference attention (block 3)	17.54	48.72	0.97
w/o global attention (blocks 12–17)	9.23	9.89	4.77
Naive stacking	20.57	57.17	4.77

A.3 Dataset Curation and Splits

Since the ECG-gated cine volumes are used as reconstruction targets, pre-existing inter-slice misalignment caused by inconsistent breath-holds would provide spatially inconsistent supervision and penalize correct reconstructions at evaluation. We therefore curate all subjects from the five public datasets. We first automatically measure the in-plane shift of each short-axis slice relative to its neighbors, using both the segmented heart position and image registration, and retain subjects whose largest inter-slice shift is at most 5 mm. We then visually inspect every subject on long-axis (LAX) views to confirm the decision. Curation is applied to the training, validation, and test sets alike, yielding the 898 subjects used in our experiments (Table A3). Fig. A4 shows representative retained and discarded subjects.

Table A3: Subjects retained after data curation.

	ACDC	M&Ms	CMRxRecon 2023	CMRxRecon 2024	CMRxRecon 2025	Total
Retained / total	116/148	189/341	141/195	211/294	241/359	898/1337

The curated data are split into 628/90/180 subjects for training, validation, and testing, respectively, while approximately balancing dataset source, scanner/site, and pathology distributions across splits. To further evaluate generalization to unseen acquisition domains, all Canon scans and all Philips scans from CMRxRecon 2025 Center012 are excluded from training and used only for validation and testing. For preprocessing, cine images are resampled to an in-plane resolution of 1.4 mm while retaining the native slice count and through-plane spacing, and each cine is temporally represented using $T = 12$ cardiac phases.

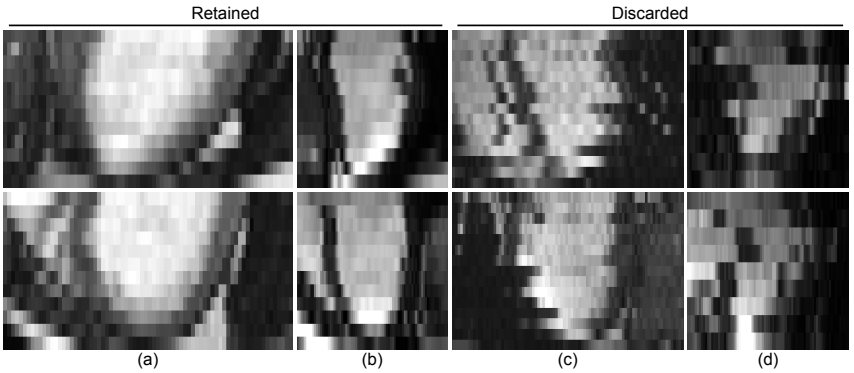

Figure A4: Example subjects (a), (b) retained and (c), (d) discarded during data curation, each shown in two orthogonal long-axis views (top and bottom). Retained subjects show a smooth ventricular contour, whereas discarded subjects show clear inter-slice misalignment caused by inconsistent breath-holds.

A.4 Detailed Rhythm Simulation

For test-time simulation, each short-axis slice is treated as a short real-time cine series with its own cardiac beats and respiratory phase. Respiratory motion evolves continuously at approximately one breathing cycle per five cardiac cycles (Fürnrohr & Heckel, 2026).

For atrial fibrillation (AF), we introduce beat-to-beat cycle-length variation following Fürnrohr & Heckel (2026). The duration of each beat, normalized by the subject-specific cardiac cycle length, is independently sampled from $\mathcal{N}(1, 0.25^2)$, consistent with the R–R variability reported in AF (Markl et al., 2015). Within each beat, the cine progresses at its original temporal rate. A beat longer than the nominal cycle first completes the cardiac cycle and then remains in the end-diastolic state for the residual duration, reflecting that cycle-length prolongation predominantly extends the diastolic rest period (Chung et al., 2004). Conversely, a short beat is terminated before completion, such that the subsequent beat begins from a partially filled ventricular state, approximating the reduced filling and ejection associated with short R–R intervals (Gosselink et al., 1995).

For heart rate variability (HRV), we simulate milder physiological beat-to-beat variation, motivated by the intrinsic variability of normal sinus rhythm (Malik et al., 1996). Following Fürnrohr & Heckel (2026), the normalized duration of each beat is sampled from $\mathcal{N}(1, 0.15^2)$, and the entire cardiac cycle is temporally scaled to match the sampled duration. In both HRV and AF simulations, the cardiac anatomy and motion are inherited from the subject-specific gated cine; only their temporal progression is modified. Intermediate cardiac states are obtained by linear interpolation between adjacent cine phases.

Each frame of the reference slice defines a target cardiac state, with the corresponding breathing-free volume at end expiration used as ground truth. For single-frame methods, including ours, one frame is provided per slice: the reference slice is fixed to the target frame, while a frame is independently sampled from each remaining slice. Dense-temporal baselines instead receive the complete temporal series of every slice.

A.5 Prospective Real-Time CMR Acquisition

The in-house prospective dataset consists of one healthy volunteer and one patient with atrial fibrillation (AF). Both participants provided written informed consent under an Institutional Review Board (IRB)-approved, HIPAA-compliant protocol. Free-breathing, ungated 2D balanced steady-state free precession (bSSFP) images were acquired on a 1.5T MAGNETOM Sola scanner (Siemens Healthineers) using a spiral trajectory with 48 interleaves for fully sampled k-space. For each subject, 10 short-axis slices were acquired sequentially to provide full left ventricular coverage. The acquisition used a 128×128 in-plane matrix with a spatial resolution of 2.3×2.3 mm², an 8-mm slice thickness, and a 2-mm interslice gap. Each slice was acquired over 4.8 s with a temporal resolution of 25 ms per frame. Neither subject was used during model training.

A.6 Evaluation and Implementation Details

Evaluation protocol. To remove global rigid coordinate-frame ambiguity, each reconstructed cardiac phase is aligned to its corresponding target using 6-DOF rigid registration before scoring. All image-based metrics are computed within the heart ROI. For methods that explicitly deconvolve the imaging point-spread function (PSF), we reconvolve the reconstructed volume with a matched subject-specific PSF before evaluation to ensure comparable spatial resolution across methods. NeSVoR and NiftyMIC output volumes on an arbitrary intensity scale, so we map their intensities to those of the input stack by matching the 0.5th and 99.9th percentiles. CiNeVol, being a continuous representation, produces values outside its fitting region; these are set to zero outside the heart ROI, as for the other methods. All methods are evaluated on the same set of slices, corresponding to those used in the single-frame input for the first target and also contained in the dense-temporal inputs.

Image quality and respiratory motion. We use normalized cross-correlation (NCC) as the primary image-quality metric because it is insensitive to method-dependent intensity scaling and offset, and additionally report peak signal-to-noise ratio (PSNR) and per-slice structural similarity index (SSIM). Respiratory motion recovery is evaluated using the through-plane endpoint error (EPE), defined as the mean absolute difference between the recovered and simulated through-plane respiratory shifts after removing the per-subject mean error. Dangi et al. (2018) estimate only in-plane motion and are therefore excluded from the through-plane EPE evaluation.

Anatomical and functional accuracy. We apply the same pretrained cardiac segmentation model (Full et al., 2021) to each reconstructed cine and its corresponding target. Anatomical accuracy is quantified using Dice and the 95th-percentile Hausdorff distance (HD95), averaged over the left ventricle (LV), myocardium (MYO), and right ventricle (RV) across all cardiac phases. Functional accuracy is assessed using the mean absolute errors in ejection fraction (EF) and left-ventricular mass (LVM). We additionally report the LV volume–time curve (VTC) error, defined as the mean absolute LV volume error over the full cardiac cycle normalized by the ground-truth end-diastolic volume (EDV).

Runtime. Runtime is measured as the wall-clock time required to reconstruct the full 3D cine for one subject. GPU-based methods (Ours, NeSVoR, Dangi et al. (2018), and CiNeVol) are evaluated on four NVIDIA L40S 48 GB GPUs, whereas CPU-based methods (SVRTK, NiftyMIC, and Fetal CMR 4D) use 16 cores of an Intel Xeon Gold 6154 3.0 GHz processor. Methods that reconstruct cardiac phases independently process phases in parallel across the available GPUs or CPU cores. CiNeVol performs its joint reconstruction using data parallelism across four GPUs, while Fetal CMR 4D performs its joint optimization using 16 CPU cores.

Implementation details. We initialize Cardiac VGGT from VGGT-1B (Wang et al., 2025a) and freeze the pretrained DINOv2 image encoder (Oquab et al., 2024) while fine-tuning the remaining network. The model operates on 518×518 inputs with a 518×518 splatting grid. We train for 300 epochs with one subject per optimization step using AdamW (Loshchilov & Hutter, 2019), with a learning rate of 5×10^{-5} and weight decay of 0.05. The learning rate follows a 5% linear warm-up followed by cosine decay. During training, shared geometric and intensity augmentations are applied across all cardiac phases, together with simulated low-resolution, Gibbs ringing, and phase-encoding ghosting artifacts; respiratory augmentation follows the respiratory motion simulation in Sec. 4. We optimize the overall objective in Sec. 3.4 with $\lambda_h = \lambda_o = 0.5$ and $\lambda_s = 1000$, and select the checkpoint achieving the highest heart-region PSNR on the validation set. Training is performed on a single NVIDIA L40S 48 GB GPU.

A.6.1 Baseline Configurations

For all baselines, we follow the original implementations and default configurations as closely as possible, deviating only where required by our single-stack, free-breathing setting or to use a common output resolution. Single-frame methods (SVRTK, NiftyMIC, NeSVoR, and Dangi et al. (2018)) receive exactly the same input as ours, i.e., one frame per slice with the reference slice at the target frame, and reconstruct each target cardiac state independently. Dense-temporal methods (Fetal CMR 4D and CiNeVol) receive all frames of every slice. Where required, methods are additionally provided with the heart ROI, the nominal slice positions, and the subject-specific slice thickness. All volumetric reconstructions use an isotropic resolution of 1.4 mm.

SVRTK (Uus et al., 2020) and NiftyMIC (Ebner et al., 2020). We use rigid slice-to-volume reconstruction with default settings. For SVRTK, robust statistics are disabled, since with a single stack every slice is otherwise rejected as an outlier.

NeSVoR (Xu et al., 2023). Stack-to-stack registration is disabled since a single stack is provided, and the heart ROI is used as the sampling mask.

Dangi et al. (2018). We reimplement the method and train it with the original protocol on our training split, using left-ventricular (LV) centers from segmentation as targets. Each slice is translated in-plane so that its predicted LV center aligns with that of the reference slice.

Fetal CMR 4D (van Amerom et al., 2019). We reconstruct 12 cardiac phases with the authors' default settings. Since the original cross-slice synchronization requires spatially overlapping slices, we instead follow Åkesson et al. (2025) and detect end-diastole in each slice as the frame with the largest LV area, segmented by a pretrained network (Full et al., 2021), and assign cardiac phases relative to it with a nominal R–R interval.

CiNeVol (Vogt et al., 2026). As the official implementation is not publicly available, we reimplement it following the original paper. In place of ECG and respiratory-belt signals, we provide the exact R-peak times from our simulation, converted to cardiac states with the original phase model, and the true simulated breathing levels as respiratory states. The model is fitted to the single short-axis stack within the heart ROI.

A.7 Pseudocode for Slice-Volume Operators

Algorithm 1 gives the slice-to-volume splatting operator $\overrightarrow{\mathcal{S}}$ and the volume-to-slice sampling operator $\overleftarrow{\mathcal{Q}}$, where w denotes the trilinear interpolation weight and $\epsilon = 10^{-6}$. Coordinates are converted to voxel units of the target grid using the subject-specific slice spacing. In practice, pixels with intensity at or below $\tau = 10^{-3}$ (e.g., zero padding) are excluded from splatting and from $\mathcal{L}_{\text{obs}}$. Both operators are differentiable with respect to the predicted displacements.

Algorithm 1 Differentiable slice-volume operators.

function SPLAT($\hat{\boldsymbol{U}}, \mathcal{I}$) ▷ Eqs. (6)–(8)
  $A \leftarrow 0,\ C \leftarrow 0$
  for each slice i and pixel $\boldsymbol{p}$ with $I_i(\boldsymbol{p}) > \tau$ **do**
    $\widetilde{\boldsymbol{g}} \leftarrow \boldsymbol{g}_i(\boldsymbol{p}) + \hat{\boldsymbol{u}}_i(\boldsymbol{p})$
    if $\widetilde{\boldsymbol{g}}$ lies inside the grid **then**
      for each of the 8 voxels $\boldsymbol{q}$ neighboring $\widetilde{\boldsymbol{g}}$ **do**
        $A(\boldsymbol{q}) \leftarrow A(\boldsymbol{q}) + w(\boldsymbol{q}, \widetilde{\boldsymbol{g}})\, I_i(\boldsymbol{p})$
        $C(\boldsymbol{q}) \leftarrow C(\boldsymbol{q}) + w(\boldsymbol{q}, \widetilde{\boldsymbol{g}})$
      end for
    end if
  end for
  return $\hat{V} = A/(C + \epsilon)$
end function

function SAMPLE($\hat{\boldsymbol{U}}, V$) ▷ Eq. (9)
  for each slice i and pixel $\boldsymbol{p}$ **do**
    $\widetilde{I}_i(\boldsymbol{p}) \leftarrow \mathrm{Trilinear}\big(V, \boldsymbol{g}_i(\boldsymbol{p}) + \hat{\boldsymbol{u}}_i(\boldsymbol{p})\big)$ ▷ zero outside the grid
  end for
  return $\widetilde{\mathcal{I}}$
end function

B Related Work (Continued)

B.1 Free-breathing and Ungated Cardiac MRI

Conventional cine CMR typically uses segmented acquisitions over multiple heartbeats, with ECG synchronization to resolve cardiac motion and repeated breath-holds to suppress respiratory mo-

tion (Kramer et al., 2020). While effective under regular physiology, this acquisition paradigm can be challenging for patients with limited breath-hold capacity and becomes less reliable under irregular cardiac rhythms. Free-breathing approaches therefore seek to relax respiratory control while preserving cardiac synchronization. Early methods used respiratory navigators or triggering together with ECG gating (Moghari et al., 2015; Pednekar et al., 2018), whereas self-gated techniques infer cardiac and respiratory motion directly from the acquired MR data, eliminating the need for external physiological sensors (Larson et al., 2004; Feng et al., 2016). Free-running frameworks further extend this idea to continuous acquisitions, jointly resolving cardiac and respiratory dimensions from self-gating signals (Di Sopra et al., 2019).

Importantly, removing external gating does not necessarily eliminate physiological synchronization. Most self-gated and free-running approaches still estimate cardiac or respiratory states and retrospectively assign measurements to motion-resolved bins (Feng et al., 2016; Di Sopra et al., 2019). Such reconstruction relies on measurements assigned to the same state being sufficiently consistent across cardiac cycles, an assumption that becomes less reliable under substantial beat-to-beat variation or arrhythmia. A complementary direction is real-time cine MRI, which reconstructs cardiac dynamics directly from continuously acquired data without retrospective cardiac-phase binning and can operate during free breathing without ECG gating (Zhang et al., 2010). However, real-time cine is typically reconstructed independently for each imaging plane. When a stack of slices is acquired sequentially, different slices may therefore still represent different respiratory positions and cardiac states, leaving their spatial and physiological consistency unresolved.

B.2 Slice-to-Volume Reconstruction

Slice-to-volume reconstruction (SVR) provides a framework for recovering a spatially coherent volume from such misaligned 2D observations. Early methods reconstruct the volume by iteratively aligning slices from several stacks acquired in orthogonal orientations (Rousseau et al., 2006). Later methods improve robustness with intensity matching and outlier removal (Kuklisova-Murgasova et al., 2012) and remove manual steps by automatically localizing and segmenting the fetal brain (Ebner et al., 2020). SVRTK (Uus et al., 2020) extends SVR to deformable motion of the fetal body and placenta, and NeSVoR (Xu et al., 2023) models the volume as a continuous implicit neural representation. However, these methods require iterative optimization for each subject and typically rely on multiple stacks. Learning-based methods instead predict slice poses directly with a neural network (Hou et al., 2018; Xu et al., 2022), but they predict each slice's pose separately or are designed for multiple stacks. FC-SVR (Young et al., 2024) predicts dense slice motion from a single stack and splats the slices into a volume. All these methods, however, reconstruct a single 3D anatomy from all slices, which does not hold for the beating heart.

For cardiac imaging, Dangi et al. (2018) correct the misalignment of conventional cine stacks by aligning CNN-predicted left-ventricular centers, assuming only translational misalignment. Other methods acquire a real-time series at every slice and assign each frame a cardiac state. Fetal CMR 4D (van Amerom et al., 2019) estimates the heart rate of each slice from its image series and aligns the cardiac cycles across slices, but a constant heart rate per slice does not accommodate beat-to-beat variation. CiNeVol (Vogt et al., 2026) extends NeSVoR to a dynamic implicit neural representation conditioned on cardiac and respiratory states derived from ECG and a respiration belt, and is optimized for each subject. Because its cardiac states come from the ECG, irregular rhythms would require further adaptation. Existing cardiac SVR methods therefore rely on conventional gated cine or on repeated observations of every slice with physiological synchronization. In contrast, our method reconstructs coherent cardiac volumes from free-breathing slices in a single forward pass, without gating or synchronization.

C Additional Qualitative Results

Figs. A5–A9 show full-cine reconstructions of a representative test subject from each of the five datasets. Each column is one target frame, so each row plays the cine over one cardiac cycle. In the short-axis views, CardioStitch follows the contraction and relaxation of the target; in the long-axis cuts, the slices form a continuous ventricle. The error is largest at tissue boundaries, where the reconstruction is smoother than the target.

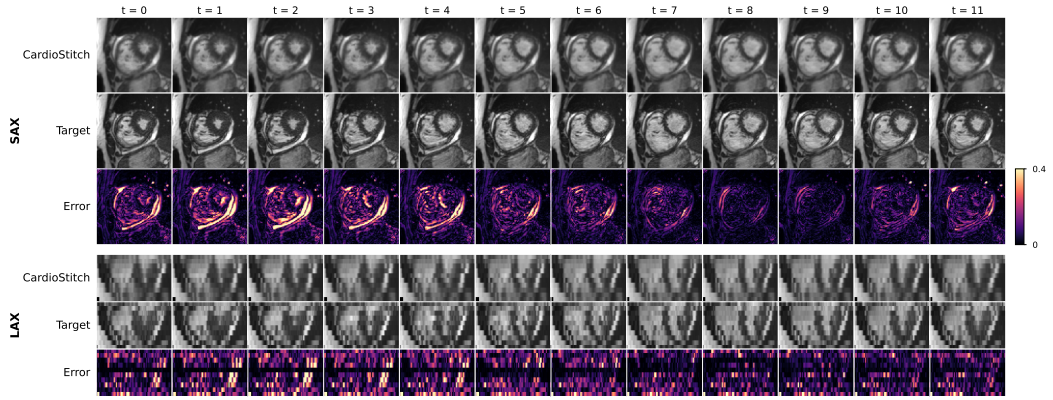

Figure A5: Full-cine reconstruction of a representative test subject from ACDC. Columns are the 12 target frames of one cardiac cycle. The top block shows a short-axis slice two slices apical to the reference slice, and the bottom block a long-axis cut through all slices. Rows: CardioStitch, target, and absolute error.

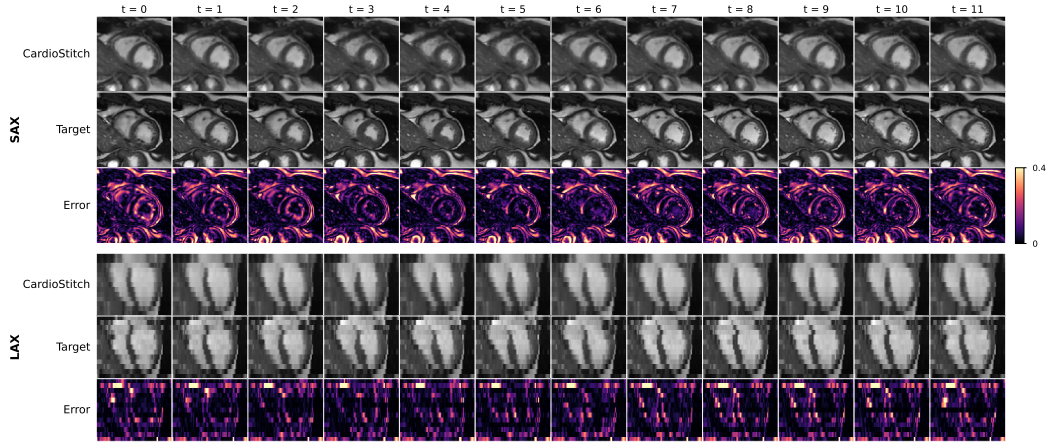

Figure A6: Same as Fig. A5, for a test subject from M&Ms.

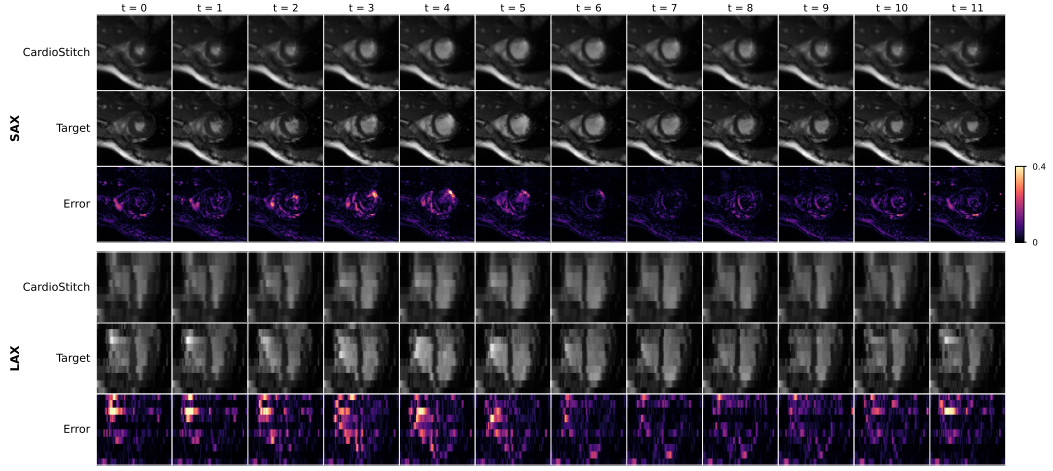

Figure A7: Same as Fig. A5, for a test subject from CMRxRecon 2023.

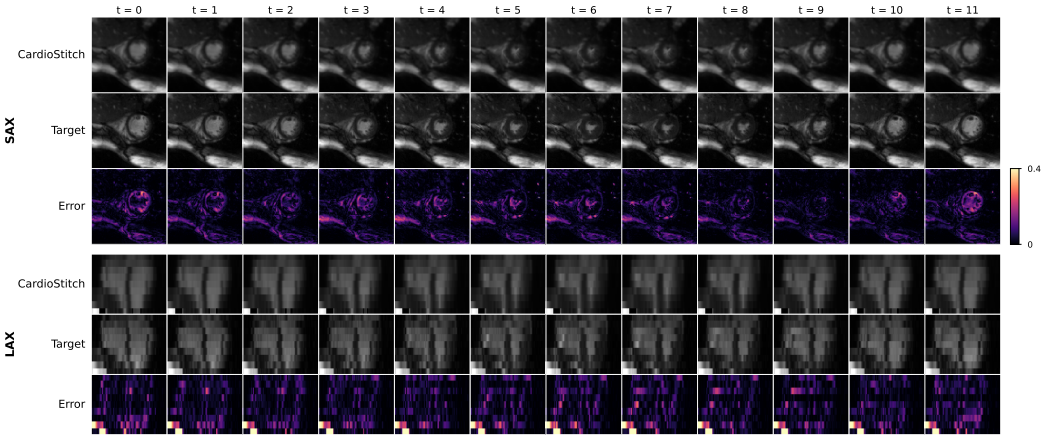

Figure A8: Same as Fig. A5, for a test subject from CMRxRecon 2024.

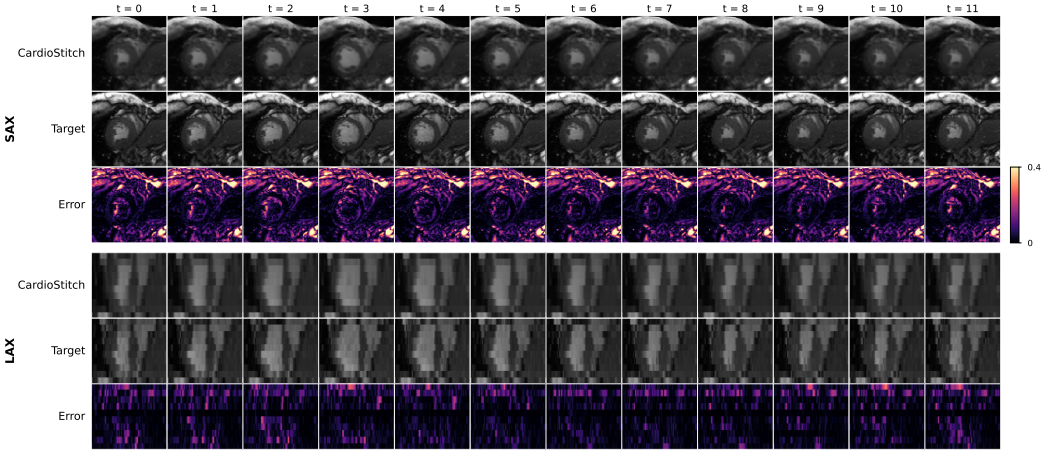

Figure A9: Same as Fig. A5, for a test subject from CMRxRecon 2025.