Focus, Align, and Diffuse: Time-Series-Aware Keyframe Diffusion for Cardiac Dynamic Synthesis from Sparse Observations

Junkai Liu[1], **Haofan Wu**[1], **Nay Aung**[2], **Joao A. C. Lima**[3], **Steffen E. Petersen**[2], **Le Zhang**[1†]

[1] School of Engineering, University of Birmingham
[2] William Harvey Research Institute, Queen Mary University London
[3] Division of Cardiology, Johns Hopkins University School of Medicine

l.zhang.16@bham.ac.uk

Project Page **GitHub**

Abstract

Dense cardiac dynamics are essential for assessing cardiac function, yet in practice clinicians often observe only extremely sparse measurements. However, recovering full dynamics from such sparsity is a highly ill-posed problem due to the lack of temporal cues. While auxiliary time-series signals like ECG provide physiological guidance, effectively exploiting them is challenging due to the cross-modal representation gap and phase misalignment. In this work, we present **CardioFAD**, a **Focus–Align–Diffuse** paradigm that reformulates sparse cardiac dynamic synthesis as a *recoverability–alignability anchor discovery* problem. Our core insight is that resolving extreme temporal ambiguity requires a compact set of anchor states that are both sufficient to recover the visual trajectory and reliable for ECG–visual alignment. We therefore learn motion-salient keyframes under two complementary principles: *trajectory sufficiency*, which preserves trajectory-critical intra-modal motion content, and *anchor alignment*, which promotes consistent inter-modal correspondence at these anchors. We implement the **Focus** and **Align** modules optimized by the two aspects. **Diffuse** then performs two-stage generation by first synthesizing keyframes and subsequently interpolating the remaining frames. Extensive experiments on cardiac MRI and echocardiography videos demonstrate superior visual quality, temporal coherence, and improved functional consistency under the extreme one-shot regime.

1 Introduction

Dynamic cardiac imaging modalities, such as cardiac MRI (CMR) and echocardiography (ECHO), are widely used to capture cardiac motion and monitor disease progression for assessing heart function [9, 27, 48]. Many clinically relevant cues are characterized by motion over the cardiac cycle [54, 71], yet acquiring sequences (3D+t CMR or 2D+t ECHO videos) is expensive and time-consuming, and clinicians often have access to only a sparse observation (e.g., a single frame or a single slice) [38, 70]. Despite progress in generating entire sequences [34, 36, 10], the problem remains underexplored, especially under extreme one-shot sparsity due to its *ill-posed* nature.

The main challenge of this task lies in the fact that vastly different spatiotemporal trajectories can explain the same limited evidence [76, 13, 32, 28, 29, 44, 74]. Given only an extremely sparse

†: Corresponding author

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

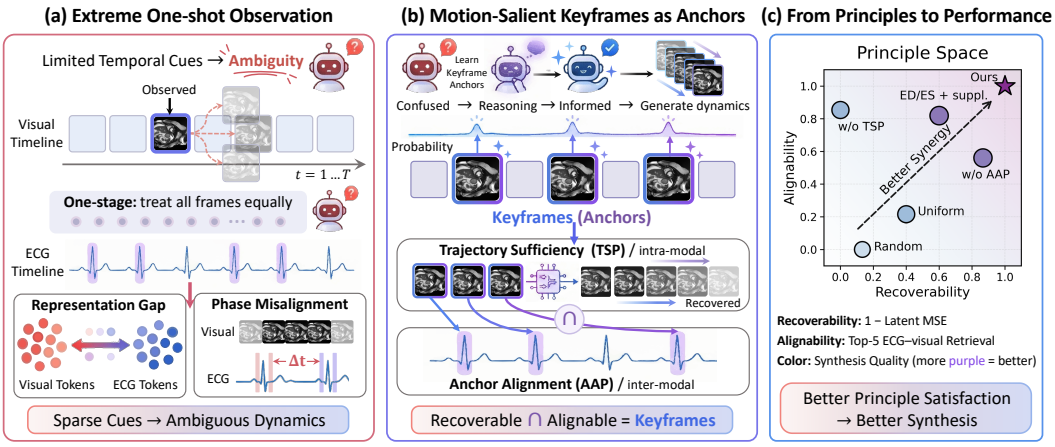


Figure 1: **(a) Challenge:** Extreme sparsity causes temporal ambiguity, while ECG suffers from cross-modal misalignments. **(b) Solution:** We reformulate the task as a recoverability–alignability anchor discovery problem. **(c) Principle-to-Performance:** Our learned keyframes satisfy both principles, yielding superior synthesis quality.

visual observation (e.g., a single frame or slice), the model lacks sufficient motion cues to anchor the global cardiac-state trajectory (Fig. 1a). This makes rapid contraction and relaxation phases highly underdetermined and prone to temporal drift or motion inconsistency [73, 53, 68]. Intuitively, not all time points are equally informative for characterizing the trajectory, thus selecting salient keyframes is critical to preserving essential cardiac motion content [66, 83, 2]. Randomly or uniformly sampled keyframes are often suboptimal, while using predefined phases such as end-diastole (ED) and end-systole (ES) as anchors is too coarse to capture richer dynamics beyond ED/ES across diverse pathologies and acquisition protocols. Meanwhile, although ECG is a natural source of auxiliary cues, translating it into reliable visual guidance is non-trivial. ECG provides phase-related electrophysiological cues but does not specify anatomy or appearance, and its correspondence to imaging states can vary across subjects, introducing a cross-modal representation gap and phase misalignment [50, 60, 11, 34] (Fig. 1a). In this context, existing one-stage conditional diffusion models treat each frame equally, and may ignore the weak conditioning, over-smooth salient motion, or hallucinate temporally inconsistent dynamics [20, 45, 7, 26]. Hence, an effective solution should learn motion-salient anchors and align ECG with visual dynamics in a temporally aware manner, so that the ECG condition can reliably constrain generation.

Motivated by the above analysis, we ask the central question: *How can we synthesize dense cardiac dynamics from an extremely sparse observation while effectively leveraging auxiliary time-series signals?* Our **key insight** is to reformulate this ill-posed task as a *recoverability–alignability anchor discovery* problem. Rather than treating all frames equally or relying on stronger global conditioning, the model should identify a compact set of motion-salient states that are both (i) sufficient to recover the intra-modal cardiac trajectory and (ii) reliable for ECG–visual alignment (Fig. 1b). Such recoverable-and-alignable anchors reduce temporal ambiguity and turn auxiliary time-series signals into more effective guidance for dense dynamic synthesis.

To exploit this recoverability–alignability anchor discovery view, we introduce **CardioFAD**, a **Focus–Align–Diffuse** paradigm for time-series-conditioned cardiac dynamic synthesis from sparse observations, as shown in Fig. 2. Instead of treating keyframes as predefined landmarks or merely adding stronger global conditioning, CardioFAD learns anchors that satisfy two principle-level criteria: (i) **trajectory sufficiency**, where informative keyframes should preserve enough intra-modal evidence to reconstruct the full cardiac trajectory; and (ii) **anchor alignment**, where these keyframes should provide reliable ECG–visual correspondence at motion-salient phases. Concretely, **Focus** learns sparse yet informative keyframes from input ECG in a self-discovered manner by enforcing the *Trajectory Sufficiency Principle (TSP)*, encouraging the model to allocate capacity to a distinct anchor set that best explains the remaining dynamics. **Align** implements the *Anchor Alignment Principle (AAP)* with temporally aware, keyframe-weighted ECG-visual alignment, thus turning ECG into reliable guidance. Finally, **Diffuse** performs two-stage diffusion by first generating anchor keyframes and then interpolating the remaining frames, yielding temporally coherent dynamics and improved functional fidelity under extreme sparsity (Fig. 1c). Our contributions can be summarized as follows:

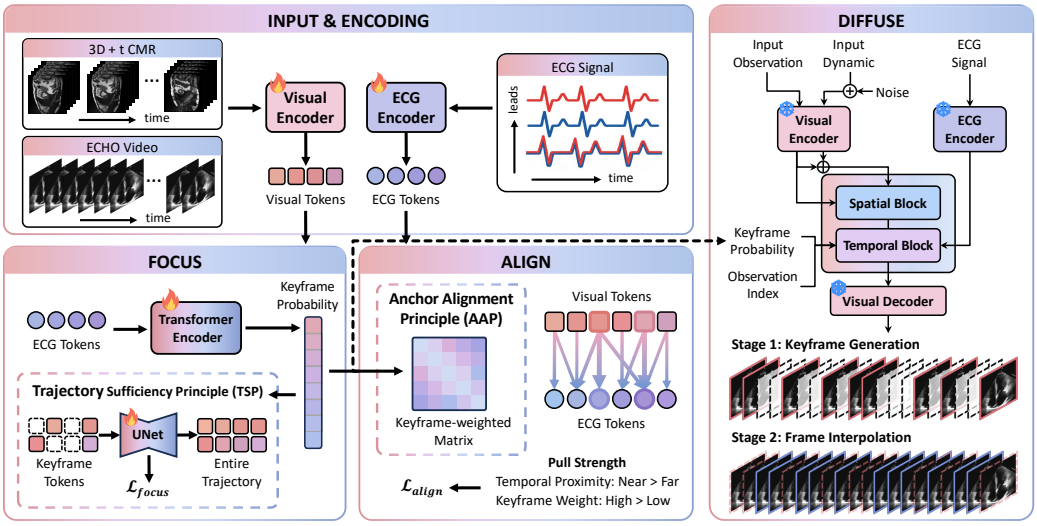


Figure 2: **Overview of CardioFAD.** Our pipeline contains three stages: **(i) Focus:** A tokenizer predicts keyframe probabilities, optimized by the TSP to ensure the selected anchors can reconstruct full intra-modal dynamics. **(ii) Align:** Guided by the AAP, a keyframe-weighted matrix synchronizes visual and ECG latents, establishing keyframes as cross-modal anchors. **(iii) Diffuse:** A video DiT performs two-stage synthesis, first generating the keyframes, then interpolating the remaining frames.

- We reformulate one-shot cardiac dynamic synthesis as a *recoverability–alignability anchor discovery* problem, and propose CardioFAD, a keyframe-to-sequence diffusion framework that learns anchors sufficient for visual trajectory recovery and reliable for ECG–visual alignment.
- We instantiate this reformulation with a novel trajectory-aware tokenizer guided by TSP and AAP, enabling self-emergent keyframe discovery and keyframe-weighted cross-modal alignment under a unified principle-driven design.
- Comprehensive experiments on both 3D+t CMR and 2D+t ECHO benchmarks demonstrate our superior performance in generative quality and cardiac functional consistency under the extreme one-shot regime. Furthermore, principle-targeted diagnostics demonstrate that our learned anchors jointly satisfy trajectory recoverability and ECG–visual alignability, justifying our reformulation of sparse cardiac synthesis as a recoverability–alignability anchor discovery problem.

2 Related Work

Cardiac Dynamic Synthesis from Medical Imaging. Cardiac dynamics synthesis reconstructs a full cardiac cycle from partial observations, enabling spatiotemporal and controllable completion [1, 37, 53]. Existing CMR studies under limited acquisition address temporal interpolation [29, 73], motion reconstruction [5, 31], super-resolution [55, 82], and conditional generation [38, 44]. For ECHO, video synthesis methods have been studied under diverse conditioning settings [57, 15, 81, 47, 74]. Despite these advances, existing methods typically rely on richer visual evidence, modality-specific assumptions, or global conditioning signals, making them less suited to one-shot cardiac synthesis. CardioFAD instead learns motion-salient keyframes that are both trajectory-sufficient and ECG-alignable, and uses them as anchors for two-stage generation across 3D+t CMR and 2D+t ECHO.

Diffusion-based Video Generation. Diffusion models have become a dominant paradigm for video generation [59, 51, 43, 23, 77]. Controllable generation further incorporates constraints such as reference frames, motion guidance, and temporally varying control signals [67, 69, 45, 75, 72]. Recent works show that conditioning on structured temporal information can substantially improve temporal coherence and enable targeted manipulation [16, 39, 17, 3, 78, 79]. Despite these advancements, existing conditioning does not explicitly handle the extreme observation sparsity in our task context. Our two-stage strategy first generates keyframes and then interpolates remaining frames, enabling effective guidance from sparse observations and physiological signals.

Keyframe Selection for Video Processing. Keyframe selection aims to identify a compact set of frames that best summarizes video content [80, 83, 62, 19]. It is widely used for video understanding [62, 33, 40, 8, 61], and for efficient modeling by focusing computation on informative

locations [6, 83, 65, 22, 30]. For video generation, keyframes can serve as fundamental anchors that capture salient states, enabling coarse-to-fine synthesis [2, 24, 66, 18, 4]. Nevertheless, existing keyframe-based methods typically assume single-modality inputs and rely on predefined or externally supervised anchors, with limited consideration of cross-modal alignment. In contrast, CardioFAD learns self-emergent keyframes from time-series guidance, selecting anchors that are both intra-modally trajectory-sufficient and inter-modally alignable for two-stage cardiac synthesis.

3 Methodology

3.1 Problem Formulation

We aim to reconstruct a complete cardiac cycle from sparse visual observations, guided by synchronized time-series (ECG). We validate our approach on two modalities: **(i) 3D+t CMR synthesis**, recovering a full 3D+t sequence from a single 2D slice or 3D volume; and **(ii) 2D+t ECHO video synthesis**, generating a complete 2D+t video from a single frame. CMR uses paired patient-specific ECG, while ECHO uses a shared patient-independent phase template for temporal guidance.

3.2 Focus: Cardiac Dynamic Keyframe Selection

Given a cardiac sequence and an ECG signal, our visual and ECG encoders first extract trajectory tokens $\mathbf{z}_t$ and time-series embeddings $\mathbf{h}_t$, respectively. Our tokenization scheme is elaborately designed to map these heterogeneous inputs into a shared temporal length T for subsequent processing.

Keyframe selection. Our keyframe selector operates in the cardiac state space, which yields learnable keyframes directly from the input ECG, empowering our tokenizer to perceive clinically meaningful phases in a self-emergent manner. Let T denote the sequence length in the latent space. A lightweight transformer [64] f_ϕ consumes the ECG tokens and outputs scalar scores ℓ_t as

$$\ell_{1:T} = f_\phi(\mathbf{h}_{1:T}), \qquad q_t = \frac{\exp(\ell_t)}{\sum_{k=1}^{T} \exp(\ell_k)}, \tag{1}$$

where q_t is the keyframe probability at time t. We produce a binary selection mask $m_t \in \{0, 1\}$ via straight-through Gumbel-Softmax top-K [25]. This yields a keyframe set $\mathcal{K} = \{t \mid m_t = 1\}$ of size K. To ensure that keyframes are informative enough, we adopt the following principle:

Trajectory Sufficiency Principle (TSP). *The selected keyframe set should be sufficient to recover the entire cardiac trajectory.*

To enforce this, we require that the model should reconstruct the entire input sequence with minimal error, making keyframes a maximally informative and compact summary of intra-modal dynamics. Specifically, we mask out non-keyframe latents by setting $\tilde{\mathbf{z}}_t = m_t \mathbf{z}_t$. A 3D UNet G_ψ then attempts to reconstruct the complete latent sequence from this sparse input as $\hat{\mathbf{z}}_{1:T} = G_\psi(\tilde{\mathbf{z}}_{1:T})$. The reconstruction loss is computed exclusively on the non-keyframe steps:

$$\mathcal{L}_{focus} = \frac{1}{T - K} \sum_{t \notin \mathcal{K}} \|\hat{\mathbf{z}}_t - \mathbf{z}_t\|_1. \tag{2}$$

This principle aligns with the intuition that the learned keyframes should encapsulate the most informative and predictive dynamics of the visual modality.

3.3 Align: Cross-Modal Cardiac Trajectory Alignment

To bridge the representation gap and mitigate phase misalignment, we synchronize the visual trajectory with the ECG signals, explicitly transforming the time-series condition into reliable guidance.

Cross-modal contrastive learning. We establish a temporally aware alignment matrix A to encode correspondences between the modalities based on cardiac phases. With both modalities mapped to a shared temporal length T, $d_{\text{time}}(k, t)$ denotes the circular distance between normalized cardiac phases, accounting for cycle wrap-around:

$$A_{k,t} = \frac{\exp\big(-d_{\text{time}}(k,t)^2/(2\sigma^2)\big)}{\sum_{i,j} \exp\big(-d_{\text{time}}(i,j)^2/(2\sigma^2)\big)}, \tag{3}$$

where σ controls the temporal neighborhood width. Using A as a soft alignment target, we compute a cross-modal contrastive loss. For the ECG-to-Visual direction, the target distribution p and the predicted similarity distribution π are formulated as:

$$p_{t,k}^{\mathrm{E}\to\mathrm{V}} = \frac{A_{k,t}}{\sum_j A_{j,t}}, \qquad \pi_{t,k}^{\mathrm{E}\to\mathrm{V}} = \frac{\exp(S_{k,t}/\tau)}{\sum_j \exp(S_{j,t}/\tau)}, \tag{4}$$

where $S_{k,t}$ is the cosine similarity between the mean-pooled ECG and visual tokens, and τ is a temperature parameter. The alignment loss for this direction is defined as:

$$\mathcal{L}_{\mathrm{align}}^{\mathrm{E}\to\mathrm{V}} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{T} p_{t,k}^{\mathrm{E}\to\mathrm{V}} \log \pi_{t,k}^{\mathrm{E}\to\mathrm{V}}. \tag{5}$$

Symmetrically, we define a visual-to-ECG loss $\mathcal{L}_{align}^{V\to E}$ by normalizing A along the other axis, yielding the base cross-modal alignment objective as $\mathcal{L}_{\mathrm{align}} = \frac{1}{2}\Big(\mathcal{L}_{\mathrm{align}}^{\mathrm{E}\to\mathrm{V}} + \mathcal{L}_{\mathrm{align}}^{\mathrm{V}\to\mathrm{E}}\Big)$.

Keyframe-weighted alignment. In the context of time-series-guided generation, we argue that keyframes play the most significant role in explaining the cross-modal relationship. To this end, we introduce the following principle:

Anchor Alignment Principle (AAP). *Keyframes serve as cross-modal anchors that dominate alignment supervision.*

To implement AAP, we assign higher alignment responsibility to keyframes by introducing a symmetric keyframe importance term based on the probabilities q in Eqn. 1, yielding $W_{k,t} = \frac{q_k+q_t}{\sum_{i,j}(q_i+q_j)}$. We then incorporate this term to reformulate the base alignment matrix $A_{k,t}$ (Eqn. 3) into:

$$\tilde{A}_{k,t} = \frac{\exp\big(-d_{\mathrm{time}}(k,t)^2/(2\sigma^2)\big)\, W_{k,t}}{\sum_{i,j}\exp\big(-d_{\mathrm{time}}(i,j)^2/(2\sigma^2)\big)\, W_{i,j}}. \tag{6}$$

Because $W_{k,t}$ scales with the keyframe probabilities, the dynamically salient keyframes automatically receive higher weights in $\tilde{A}$. This ensures that temporally synchronized and keyframe-like moments contribute more to the alignment. We replace the base matrix $A_{k,t}$ with the keyframe-weighted matrix $\tilde{A}_{k,t}$ in Eqn. 4 to compute the target distribution p, seamlessly enforcing the AAP in the overall loss $\mathcal{L}_{align}$. Ultimately, the overall loss of the tokenizer is the weighted sum of the reconstruction loss and the two principle-guided auxiliary objectives:

$$\mathcal{L}_{tok} = \mathcal{L}_{\mathrm{rec}} + \lambda_1\mathcal{L}_{\mathrm{focus}} + \lambda_2\mathcal{L}_{\mathrm{align}}. \tag{7}$$

3.4 Diffuse: Two-Stage Keyframe-to-Sequence Generation

Conditional video DiT. We adopt a modified video DiT [46, 43] as our generative backbone. It comprises a stack of alternating spatial and temporal transformer blocks. More specifically, the spatial blocks attend to spatiotemporal patches and receive global conditioning from the sparse visual observation's latent via AdaLN [51], while temporal blocks attend along the cardiac phase axis, receiving per-frame conditioning from the encoded ECG latents $\mathbf{h}_t$, the observed frame index, and the keyframe probabilities q_t via cross-attention.

Unified masked diffusion. Rather than using separate models, we formulate both keyframe generation and sequence interpolation as a unified masked diffusion task in the latent space. Given the keyframe mask $m \in \{0,1\}^T$ predicted by **Focus** (Sec. 3.2), we randomly sample a training mode at each iteration: (A) *Keyframe generation* uses m to activate the diffusion loss on keyframe positions, while (B) *Frame interpolation* uses $1-m$ to activate the loss on non-keyframe positions. During training, we sample mode $s \in \{\mathrm{A},\mathrm{B}\}$ with probability $p(s)$ and optimize a weighted objective $\mathcal{L}_{diff} = \lambda_s\mathcal{L}_s$ to balance anchor learning and trajectory completion. During inference, we first sample keyframe latents (mode A) and subsequently generate non-keyframes conditioned on these anchors (mode B). Further details are provided in Appendix B.2.

Table 1: **Quantitative results for 3D+t CMR synthesis on UKBB dataset**.

Methods	Volume2Seq					Slice2Seq				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
VideoComposer [67]	28.15	0.843	0.107	128.76	1762.34	21.86	0.497	0.289	314.85	4574.61
MotionCtrl [69]	28.46	0.825	0.104	119.52	1334.87	21.25	0.488	0.294	365.97	4914.27
Tora [78]	29.74	0.854	0.092	97.30	1025.57	23.10	0.539	0.251	267.92	3109.75
MotionPro [79]	27.88	0.817	0.116	145.39	1887.96	22.09	0.501	0.269	292.17	3756.61
CardioFAD (One-stage)	30.51	0.861	0.082	82.48	1007.63	23.75	0.559	0.231	211.49	2785.30
CardioFAD (Ours)	**32.41**	**0.885**	**0.064**	**59.96**	**736.01**	**25.60**	**0.587**	**0.198**	**163.12**	**1847.73**

Table 2: **Quantitative results for 2D+t ECHO synthesis on three datasets**.

Methods	EchoNet-Dynamic					EchoNet-Pediatric (A4C)					EchoNet-Pediatric (PSAX)				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
General Video Generation															
VideoComposer [67]	22.78	0.539	0.109	48.25	284.16	27.49	0.633	0.089	34.74	346.31	27.26	0.621	0.084	42.19	389.07
Tora [78]	23.90	0.548	0.102	52.06	241.96	27.86	0.638	0.082	28.94	361.09	28.37	0.638	0.081	35.52	347.21
ECHO Video Generation															
HeartBeat [81]	21.24	0.527	0.122	58.97	376.51	26.71	0.629	0.088	37.56	415.71	27.65	0.630	0.079	45.02	407.26
EchoNet-Synthetic [58]	23.58	0.540	0.107	49.87	267.35	26.08	0.624	0.092	39.93	408.12	26.23	0.612	0.090	48.82	432.50
ECHOPulse [34]	25.63	0.568	0.081	37.05	196.28	28.14	0.641	0.080	29.63	357.74	28.50	0.639	0.080	37.57	367.46
CardioFAD (One-stage)	25.14	0.556	0.089	46.82	232.17	28.32	0.643	0.078	29.23	339.15	28.77	0.640	0.079	39.17	381.06
CardioFAD (Ours)	**27.73**	**0.596**	**0.067**	**26.92**	**137.70**	**29.48**	**0.657**	**0.073**	**23.21**	**286.29**	**29.59**	**0.645**	**0.071**	**27.12**	**293.86**

4 Experiments

4.1 Experimental Settings

Datasets. (i) 3D+t CMR synthesis. We use the UK Biobank (UKBB) [52] dataset with paired short-axis CMR sequences and ECG recordings, which contains 4,379 samples with 50 frames per subject. **(ii) ECHO video synthesis.** We evaluate on EchoNet-Dynamic [49] and EchoNet-Pediatric [56]. For the latter, we treat A4C and PSAX as two separate datasets. For ECHO datasets without ECG, we use a patient-agnostic canonical phase template as a weak pseudo-ECG surrogate (see Appendix B.3). This template is shared across samples and does not use test-time ED/ES annotations, contours, volume tracings, or any patient-specific temporal labels. Perturbation results in Appendix C.9 further show that CardioFAD mainly relies on coarse phase cues rather than precise waveform details.

Implementation Details. We adopt LeanVAE [12] as the visual tokenizer and a lightweight ViT [14] as the ECG encoder. We set the number of keyframes to $K = 10/8$. The LeanVAE is trained jointly with the keyframe selector and alignment modules, which is justified in Appendix C.1. For generation, we adopt rectified flow [35, 41] as the training objective. We set the mode sampling probabilities to $p(A) = p(B) = 0.5$, and the objective weights to $\lambda_A = 1.5$ and $\lambda_B = 0.5$. Please refer to Appendix B.1 for more training details and hyperparameter settings.

Evaluation Metrics. We use five metrics for comparisons: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), Fréchet Inception Distance (FID) [21] for visual quality, and Fréchet Video Distance (FVD) [63] for temporal consistency. Beyond these generative metrics, we assess the cardiac functional consistency of synthesized ECHO videos via ejection fraction (EF) estimation. Following [57, 58], we train an EF regressor on training videos, and then apply it to generated videos to obtain predicted EF values (details in Appendix B.4). We report R^2, MAE, and RMSE against ground-truth EF.

Comparison Methods. Since no prior method matches our extreme-sparsity CMR setting, we compare with representative controllable video generation baselines [67, 69, 78, 79], and include ECHO-specific baselines [58, 34, 81] for ECHO synthesis. All baselines are re-trained under the same data preprocessing, observation protocol, and matched training setups. To better isolate the effect of our design, we also report a strong internal one-stage baseline, ***CardioFAD (One-stage)***, that shares the same tokenizer, backbone, and training setup as CardioFAD, but removes the proposed keyframe anchoring and two-stage generation. This controlled baseline provides the **fairest comparison** for attributing gains to the Focus–Align–Diffuse design rather than architecture or training differences.

4.2 Comparison with State-of-the-Art

We consider three *extreme-sparsity* settings: **Volume2Seq** and **Slice2Seq** for 3D+t CMR, and **Frame2Video** for 2D+t ECHO, each reconstructing the full sequence from a single observed vol-

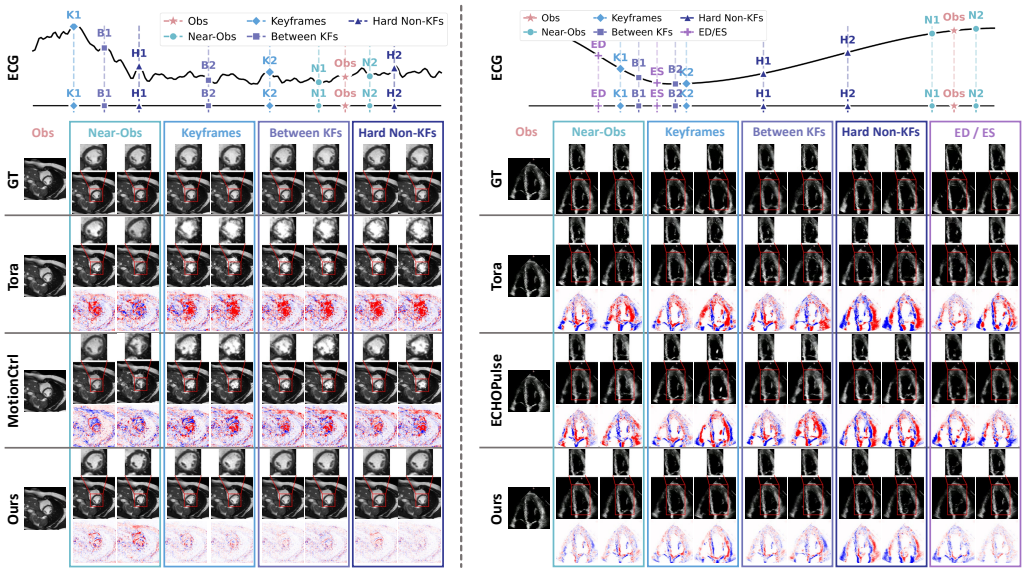


Figure 3: **Qualitative comparisons on CMR (left) and ECHO (right)**. Frames are grouped by temporal roles, with top ECG curves providing physiological context. Zoomed-in crops and difference heatmaps are provided. Compared to baselines, CardioFAD achieves higher visual fidelity and temporal consistency, avoiding phase drift particularly at challenging positions.

Table 3: **Cardiac functional consistency** evaluated by EF estimation on EchoNet-Dynamic.

Methods	R^2↑	MAE↓	RMSE↓
General Video Generation			
VideoComposer [67]	0.57	6.03	8.49
Tora [78]	0.55	6.10	8.76
ECHO Video Generation			
HeartBeat [81]	0.57	6.08	8.53
EchoNet-Synthetic [58]	0.71	5.07	7.09
ECHOPulse [34]	0.78	4.62	6.08
CardioFAD (One-stage)	0.69	4.98	7.19
CardioFAD (Ours)	**0.82**	**4.17**	**5.59**

Table 4: **Principle-guided keyframe learning ablation** on EchoNet-Dynamic.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
Keyframe Learning Principle					
w/o TSP	25.98	0.570	0.082	41.83	203.37
w/o AAP	26.74	0.581	0.079	33.96	180.29
w/o TSP & w/o AAP	25.58	0.566	0.085	44.65	239.81
Keyframe Selection Strategy					
w/ uniform keyframes	25.94	0.572	0.081	40.41	188.36
w/ random keyframes	25.17	0.554	0.091	45.59	251.03
w/ GT ED/ES	26.20	0.572	0.083	42.76	202.41
w/ GT ED/ES + Supplement	26.83	0.587	0.074	34.53	184.97
CardioFAD	**27.73**	**0.596**	**0.067**	**26.92**	**137.70**

ume/slice/frame. One-shot observation corresponds to $B{=}1$, obtained by randomly sampling one visual observation, where B denotes the number of available visual observations.

Quantitative Results. As shown in Tables 1 and 2, the internal baseline *CardioFAD (One-stage)* is already competitive with baselines, and surpasses them in multiple settings, validating our strong architectural foundation. Building upon this, *CardioFAD (Ours)* yields further improvements, particularly in temporal consistency. This substantial gain directly confirms our core motivation that explicitly anchoring generation with learned keyframes via a two-stage process is essential for resolving temporal ambiguity and ensuring coherent dynamics under extreme sparsity.

Cardiac Functional Consistency. As reported in Table 3, CardioFAD achieves the best EF regression performance on EchoNet-Dynamic among all baselines, indicating improved preservation of EF-related functional dynamics in the synthesized videos.

Qualitative Results. In Fig. 3, instead of uniformly sampling frames, we uniquely organize frames by their temporal roles alongside ECG context. While baselines suffer from motion inconsistency and blurring, especially at intermediate and hard positions, CardioFAD maintains sharp, phase-consistent dynamics across the entire cardiac cycle. This visually validates that our learned motion-salient keyframes act as robust anchors, successfully regularizing the challenging non-keyframe generation.

4.3 Robustness under Extreme Sparsity

We further evaluate $B \in \{1, 2, 4, 8\}$ to assess robustness across observation budgets. In Fig. 4, CardioFAD achieves its largest gains at $B{=}1$, where visual evidence is most limited and temporal

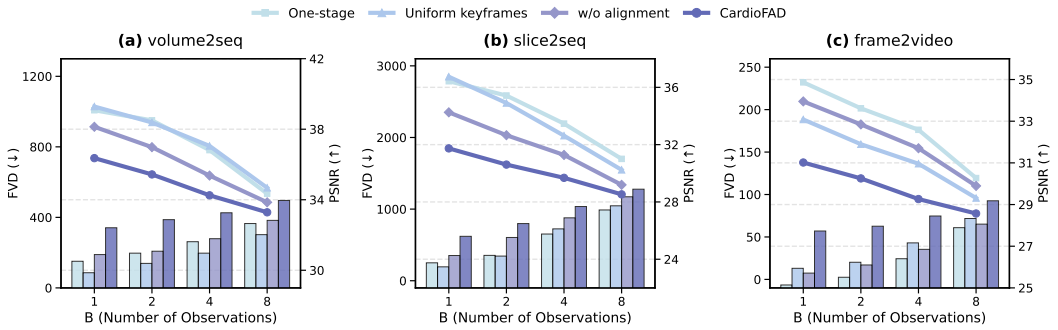


Figure 4: **Sparse observation budget**. Line charts report FVD, and bar charts report PSNR. When $B = 1$ (extremely sparse), CardioFAD exhibits its greatest advantage.

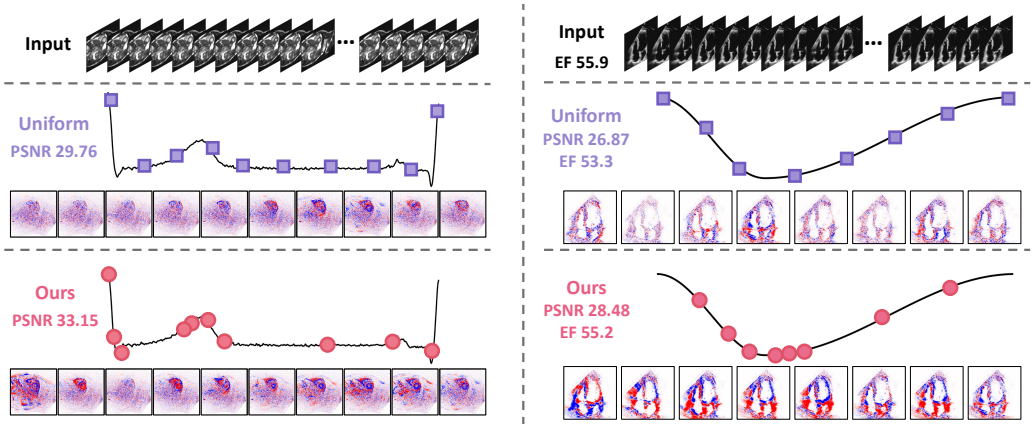


Figure 5: **Visualization of keyframe selection strategies**. Representative CMR (left) and ECHO (right) sequences with ECG overlaid. Inter-frame difference heatmaps are provided to illustrate motion magnitude. Our learned keyframes explicitly target motion-salient phases.

ambiguity is strongest. Although the gap narrows as more observations become available, CardioFAD remains competitive across all budgets, indicating robust behavior beyond the one-shot setting.

4.4 Ablation Studies

Table 4 isolates the role of principle-guided anchor discovery. Removing either TSP or AAP degrades performance, and replacing our learned anchors with random, uniform, or GT ED/ES anchors, even when ED/ES is padded to the same budget $K = 8$, consistently performs worse. These findings indicate that the gain does not come from two-stage generation alone, but from **discovering self-emergent anchors that lie in the recoverable-and-alignable paradigm**.

Keyframe Visualization. Fig. 5 further illustrates the advantage of learned keyframes over uniform sampling. Inter-frame difference heatmaps indicate that our keyframes concentrate on motion-salient phases with larger anatomical changes, providing more informative anchors for two-stage generation.

Further fine-grained ablation results and analyses of robustness to ECG perturbations are provided in Appendix C.4 and C.9, respectively. These additional ablations further separate the effects of ECG guidance, cross-modal alignment, keyframe probabilities, and reference-frame indexing, highlighting that the gains do not stem from a single conditioning source.

4.5 Principle-Targeted Diagnostics for Anchor Discovery

Beyond standard ablations, we use principle-targeted diagnostics to test the central insight of our reformulation, i.e., effective sparse cardiac synthesis requires anchors that are not only trajectory-recoverable, but also ECG–visually alignable.

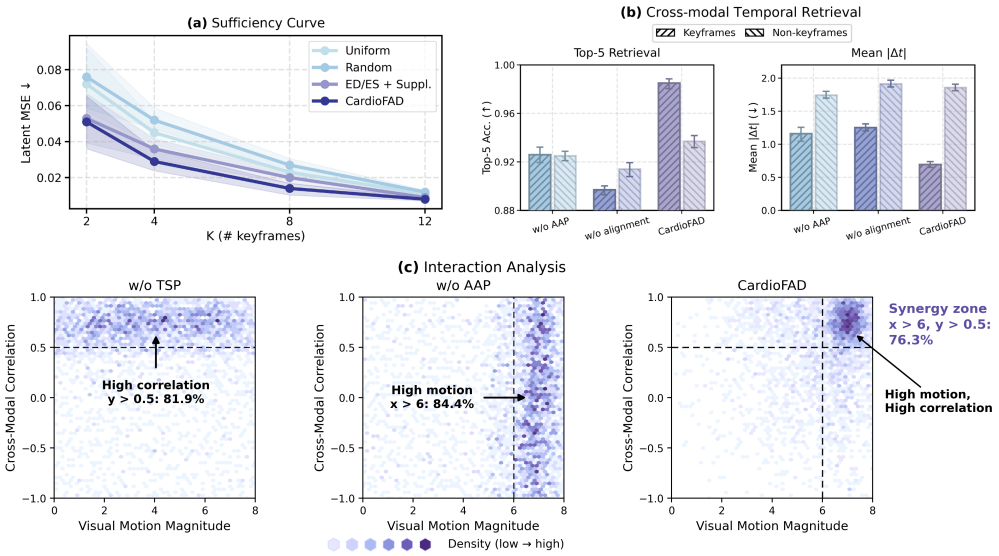


Figure 6: **(a) TSP Effectiveness:** CardioFAD achieves lower error with fewer frames, indicating higher information density. **(b) AAP Effectiveness:** Cross-modal retrieval metrics show that AAP successfully concentrates alignment capability on keyframes compared to non-keyframes. **(c) Interaction Analysis:** Visualizing keyframes in Motion-Correlation space. Removing TSP biases selection towards high-correlation frames (top), while removing AAP leads to high-motion frames (right). CardioFAD (top-right) effectively balances dynamic salience with physiological consistency.

TSP: Maximizing Information Density. *Does CardioFAD select the most informative frames?* In Fig. 6(a), we measure the reconstruction error of a trajectory. We adopt a selection-agnostic Keyframe-to-Trajectory probe trained with randomly masked keyframes. CardioFAD consistently yields the lowest latent MSE, especially at low budgets ($K = 2, 4$). This confirms that our selector prioritizes trajectory-sufficient states that maximally preserve intra-modal motion.

AAP: Aligning Cardiac Trajectories. *Does AAP concentrate ECG–visual alignment on anchors?* Fig. 6(b) reports ECG-to-visual time-step retrieval performance using the learned embeddings, measured by Top-5 accuracy and mean absolute temporal error, $|\Delta t|$. Our model achieves satisfactory alignment accuracy specifically on keyframes compared to non-keyframes. Removing AAP degrades this anchor-specific performance, verifying that AAP concentrates alignment supervision on an anchor subset, turning ECG into reliable constraints exactly where motion is most informative.

Interaction: Validating Recoverability–Alignability Anchors. *Do TSP and AAP jointly discover anchors in the recoverable-and-alignable paradigm?* Fig. 6(c) visualizes learned keyframes in the Motion Magnitude vs. Cross-Modal Correlation space, where motion magnitude measures adjacent latent displacement and cross-modal correlation measures visual–ECG cosine similarity. Effective anchors should exhibit both high motion salience and strong physiological correspondence. As shown, without TSP, selection is biased toward visually less informative but highly correlated frames; without AAP, it drifts toward high-motion frames with weaker ECG–visual correspondence. In contrast, CardioFAD concentrates keyframes in the top-right synergy zone, providing evidence that the **proposed reformulation discovers anchors that are both trajectory-recoverable and ECG–visually alignable**.

4.6 Visualization and Analysis

Cardiac Trajectory Alignment. t-SNE visualizations show that both CMR (Fig. 7(a)) and ECG (Fig. 7(b)) latents form unified, smooth manifolds with consistent phase ordering. These trends support the promotion of phase-aware cross-modal organization across modalities. The alignment matrix $\tilde{A}$ (Fig. 7(c)) exhibits sharper diagonal concentration than raw cosine similarity S, revealing that CardioFAD establishes phase-aware synchronization between the two modalities.

Keyframe Distribution. We further validate the physiological relevance of our self-discovered keyframes against GT ED/ES annotations on ECHO datasets in Fig. 8. *Qualitatively*, our keyframes

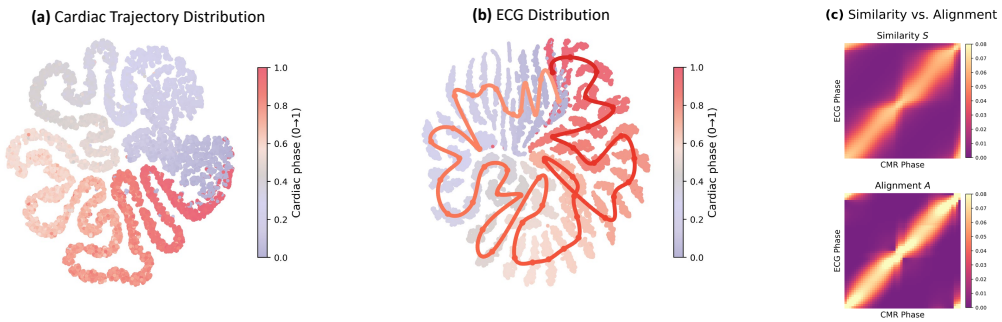


Figure 7: **Cardiac trajectory visualization**. (a) t-SNE of CMR latents. (b) t-SNE of ECG latents with a trajectory through the temporal centroids. (c) Dataset-level mean ECG-CMR similarity matrix S (top) and the corresponding alignment matrix $\tilde{A}$ (bottom).

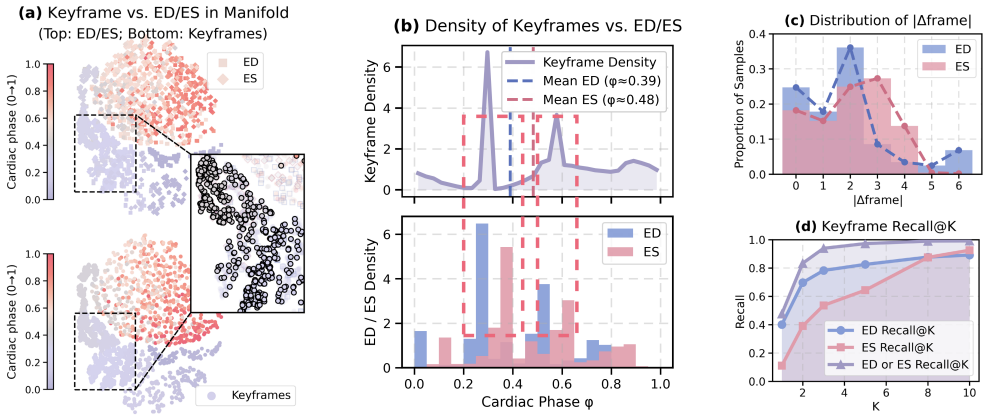


Figure 8: **Keyframe distribution visualization**. (a) GT ED/ES frames (top) and our selected keyframes (bottom) on the cardiac manifold. ED/ES-matching keyframes are outlined in black. (b) Mean keyframe phase density and ED/ES phase histograms (dashed lines: mean phases). (c) Distribution of $|\Delta\text{frame}|$ to the nearest keyframe from each ED/ES frame. (d) Keyframe Recall@K.

strongly co-localize with ED/ES in the latent manifold (Fig. 8(a)) and exhibit sharp probability peaks near clinical phases (Fig. 8(b)), demonstrating unsupervised alignment with motion-salient landmarks. *Quantitatively*, temporal proximity is impressive, with over 85% of GT ED/ES frames lying within a ± 3 frame window of a selected keyframe (Fig. 8(c)), i.e., ED/ES are typically only a few frames from at least one learned keyframe. Consequently, Recall@K rises rapidly to 99.1% at $K = 8$ (Fig. 8(d)), confirming that our method effectively captures discriminative cardiac events.

5 Conclusion

In this work, we introduce CardioFAD, a time-series-aware diffusion framework that reformulates sparse cardiac dynamic synthesis as a recoverability–alignability anchor discovery problem. Guided by TSP and AAP, CardioFAD learns self-emergent motion-salient anchors that are sufficient for visual trajectory recovery and reliable for ECG–visual alignment. These anchors enable keyframe-to-sequence generation that reduces temporal ambiguity, bridges visual anatomy with physiological signals, and achieves strong visual and functional preservation under one-shot scenarios.

Limitations. Despite the gains in Slice2Seq performance, inferring motion from 2D slices remains ambiguous and challenging. Assessing the clinical fidelity of generated CMR without explicit annotations such as segmentation labels or ED/ES phases is an open challenge we aim to address next. In addition, our current validation focuses on periodic cardiac motion, and extending the framework to non-periodic dynamics or general video generation remains to be explored.

Acknowledgments

The computations described in this research were performed using the Baskerville Tier 2 HPC service. Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham.

References

[1] N. Algethami, T. Iqbal, and I. Ullah. Generative ai for biomedical video synthesis: a review. *Artificial Intelligence Review*, 58(12):1–50, 2025.

[2] J. Bae, I. Hwang, Y.-Y. Lee, Z. Guo, J. Liu, Y. Ben-Shabat, Y. M. Kim, and M. Kapadia. Less is more: Improving motion diffusion models with sparse keyframes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11069–11078, 2025.

[3] S. Bahmani, I. Skorokhodov, G. Qian, A. Siarohin, W. Menapace, A. Tagliasacchi, D. B. Lindell, and S. Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22875–22889, 2025.

[4] A. Bigata, M. Stypułkowski, R. Mira, S. Bounareli, K. Vougioukas, Z. Landgraf, N. Drobyshev, M. Zieba, S. Petridis, and M. Pantic. Keyface: Expressive audio-driven facial animation for long sequences via keyframe interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5477–5488, 2025.

[5] N. Bubeck, S. Shit, C. Chen, C. Zhao, P. Guo, D. Yang, G. Zitzlsberger, D. Xu, B. Kainz, D. Rueckert, et al. Latent interpolation learning using diffusion models for cardiac volume reconstruction. *arXiv preprint arXiv:2508.13826*, 2025.

[6] S. Buch, A. Nagrani, A. Arnab, and C. Schmid. Flexible frame selection for efficient video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29071–29082, 2025.

[7] R. Burgert, Y. Xu, W. Xian, O. Pilarski, P. Clausen, M. He, L. Ma, Y. Deng, L. Li, M. Mousavi, M. Ryoo, P. Debevec, and N. Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13–23, 2025.

[8] X. Cao, H. Guo, J. Qian, G. Nan, C. Wang, Y. Pan, T. Hou, X. Wang, and Y. Gao. Videominer: Iteratively grounding key frames of hour-long videos via tree-based group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23773–23783, 2025.

[9] S. Chadalavada, K. Fung, E. Rauseo, A. M. Lee, M. Y. Khanji, A. Amir-Khalili, J. Paiva, H. Naderi, S. Banik, M. Chirvasa, M. T. Jensen, N. Aung, and S. E. Petersen. Myocardial strain measured by cardiac magnetic resonance predicts cardiovascular morbidity and death. *JACC*, 84(7):648–659, 2024.

[10] Y. Chen, J. Yang, D. S. Mercadier, H. Le, J. Schwitter, and P. Fua. End-to-end 4d heart mesh recovery across full-stack and sparse cardiac mri. *arXiv preprint arXiv:2509.12090*, 2025.

[11] Z. Chen, Y. Miao, L. Wang, L. Fan, D. P. Mandic, and J. Zhu. Versatile cardiovascular signal generation with a unified diffusion transformer. *Nature Machine Intelligence*, pages 1–14, 2025.

[12] Y. Cheng and F. Yuan. Leanvae: An ultra-efficient reconstruction vae for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15692–15702, 2025.

[13] S. Deb, N. Wu, F. H. Epstein, and M. Zhang. Unsupervised cardiac video translation via motion feature guided diffusion model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 648–658, 2025.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] X. Fang, Z. Ding, J. Cai, Y. Xiao, B. Liu, J. Jin, H. Wang, G. Nie, S. Huang, T. Chen, et al. Ecgflowcmr: Pretraining with ecg-generated cine cmr improves cardiac disease classification and phenotype prediction. *arXiv preprint arXiv:2601.20904*, 2026.

[16] W. Feng, T. Qi, J. Liu, M. Sun, P. Tu, T. Ma, F. Dai, S. Zhao, S. Zhou, and Q. He. I2vcontrol: Disentangled and unified video motion synthesis control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14051–14060, 2025.

[17] D. Geng, C. Herrmann, J. Hur, F. Cole, S. Zhang, T. Pfaff, T. Lopez-Guevara, Y. Aytar, M. Rubinstein, C. Sun, O. Wang, A. Owens, and D. Sun. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–12, 2025.

[18] Z. Geng, C. Han, Z. Hayder, J. Liu, M. Shah, and A. Mian. Text-guided 3d human motion generation with keyframe-based parallel skip transformer. *arXiv preprint arXiv:2405.15439*, 2024.

[19] W. Guo, Z. Chen, S. Wang, J. He, Y. Xu, J. Ye, Y. Sun, and H. Xiong. Logic-in-frames: Dynamic keyframe search via visual semantic-logical verification for long video understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[20] Y. Guo, C. Yang, A. Rao, M. Agrawala, D. Lin, and B. Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348, 2024.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[22] K. Hu, F. Gao, X. Nie, P. Zhou, S. Tran, T. Neiman, L. Wang, M. Shah, R. Hamid, B. Yin, and T. Chilimbi. M-llm based video frame selection for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13702–13712, 2025.

[23] Y. Hu, Z. Chen, and C. Luo. Lamd: Latent motion diffusion for image-conditional video generation. *International Journal of Computer Vision*, pages 1–17, 2025.

[24] H.-P. Huang, Y.-C. Su, and M.-H. Yang. Generating long-take videos via effective keyframes and guidance. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3709–3720, 2025.

[25] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

[26] H. Jeong, C.-H. P. Huang, J. C. Ye, N. J. Mitra, and D. Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7276–7287, 2025.

[27] M.-Y. Jeung, P. Germain, P. Croisille, S. E. ghannudi, C. Roy, and A. Gangi. Myocardial tagging with mr imaging: Overview of normal and pathologic findings. *RadioGraphics*, 32(5):1381–1398, 2012.

[28] S. Kalaie, A. Bulpitt, A. F. Frangi, and A. Gooya. An end-to-end deep learning generative framework for refinable shape matching and generation. *IEEE Transactions on Medical Imaging*, 44(8):3323–3344, 2025.

[29] J. Kim, H. Yoon, G. Park, K. Kim, and E. 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 (CVPR)*, pages 11353–11364, 2024.

[30] Y. Kumar. Language-guided temporal token pruning for efficient VideoLLM processing. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8924–8931, 2025.

[31] F. Laumer, L. Rubi, M. A. Matter, S. Buoso, G. Fringeli, F. Mach, F. Ruschitzka, J. M. Buhmann, and C. M. Matter. 2d echocardiography video to 3d heart shape reconstruction for clinical application. *Medical Image Analysis*, 101:103434, 2025.

[32] L. P. Le, T. Nguyen, M. A. Riegler, P. Halvorsen, and B. T. Nguyen. Multimodal missing data in healthcare: A comprehensive review and future directions. *Computer Science Review*, 56:100720, 2025.

[33] H. Lee, J. Kim, H. Kim, and Y. M. Ro. Refocus: Reinforcement-guided frame optimization for contextual understanding. *arXiv preprint arXiv:2506.01274*, 2025.

[34] Y. Li, S. Kim, Z. Wu, H. Jiang, Y. Pan, P. Jin, S. Song, Y. Shi, X. Yu, T. Yang, T. Liu, Q. Li, and X. Li. ECHOPulse: ECG controlled echocardio-gram video generation. In *International Conference on Learning Representations*, 2025.

[35] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

[36] C. Liu, X. Yuan, Z. Yu, and Y. Wang. Texdc: Text-driven disease-aware 4d cardiac cine mri images generation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 3005–3021, 2024.

[37] J. Liu, N. Aung, T. N. Arvanitis, J. A. C. Lima, S. E. Petersen, and L. Zhang. Exploiting completeness perception with diffusion transformer for unified 3d MRI synthesis. *Transactions on Machine Learning Research*, 2026.

[38] J. Liu, N. Aung, T. N. Arvanitis, S. K. Piechnik, J. A. Lima, S. E. Petersen, and L. Zhang. Sagcnet: Spatial-aware graph completion network for missing slice imputation in population cmr imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–466, 2025.

[39] L. Liu, T. Ma, B. Li, Z. Chen, J. Liu, G. Li, S. Zhou, Q. He, and X. Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14951–14961, 2025.

[40] S. Liu, C. Zhao, T. Xu, and B. Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3318–3327, 2025.

[41] X. Liu, C. Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

[42] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[43] H. Lu, G. Yang, N. Fei, Y. Huo, Z. Lu, P. Luo, and M. Ding. VDT: General-purpose video diffusion transformers via mask modeling. In *International Conference on Learning Representations*, 2024.

[44] Q. Ma, Q. Meng, M. Qiao, P. M. Matthews, D. P. O'Regan, and W. 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, 2025.

[45] W.-D. K. Ma, J. P. Lewis, and W. B. Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, 2024.

[46] X. Ma, Y. Wang, X. Chen, G. Jia, Z. Liu, Y.-F. Li, C. Chen, and Y. Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025.

[47] E. Oladokun, S. Thomas, J. Šprem, and V. Grau. Echolvfm: One-step video generation via latent flow matching for echocardiogram synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 196–206, 2026.

[48] A. M. S. Omar, M. Bansal, and P. P. Sengupta. Advances in echocardiographic imaging in heart failure with reduced and preserved ejection fraction. *Circulation Research*, 119(2):357–374, 2016.

[49] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

[50] T. Pan, Y. Sun, C. Jiang, Q. Gao, R. Sun, X. Zhang, Z. Yang, L. Han, Y. Liang, Y. Cheng, et al. Tracing the heart's pathways: Ecg representation learning from a cardiac conduction perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 890–898, 2026.

[51] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023.

[52] S. E. Petersen, P. M. Matthews, J. M. Francis, M. D. Robson, F. Zemrak, R. Boubertakh, A. A. Young, S. Hudson, P. Weale, S. Garratt, R. Collins, S. Piechnik, and S. Neubauer. Uk biobank's cardiovascular magnetic resonance protocol. *Journal of Cardiovascular Magnetic Resonance*, 18(1):8, 2016.

[53] J. Pung, G.-H. Lee, H. Huh, D. H. Yang, and H. Ha. Artificial intelligence in 4d flow mri: Review of technological aspects and clinical applications. *Artificial Intelligence in Medicine*, 171:103308, 2026.

[54] M. Qiao, K. A. McGurk, S. Wang, P. M. Matthews, D. P. O'Regan, and W. Bai. A personalized time-resolved 3d mesh generative model for unveiling normal heart dynamics. *Nature Machine Intelligence*, pages 1–12, 2025.

[55] D. Qiu, J. Shang, Y. Zhang, W. Zhang, K. K. Wong, and Z. Yi. Ccdm: Continuous-time conditional diffusion model for blind cmri super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2026.

[56] C. D. Reddy, L. Lopez, D. Ouyang, J. Y. Zou, and B. He. Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. *Journal of the American Society of Echocardiography*, 36(5):482–489, 2023.

[57] H. Reynaud, A. Gomez, P. Leeson, Q. Meng, and B. Kainz. Echoflow: A foundation model for cardiac ultrasound image and video generation. *arXiv preprint arXiv:2503.22357*, 2025.

[58] H. Reynaud, Q. Meng, M. Dombrowski, A. Ghosh, T. Day, A. Gomez, P. Leeson, and B. Kainz. Echonet-synthetic: Privacy-preserving video generation for safe medical data sharing. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 285–295, 2024.

[59] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[60] A. Selivanov, P. Müller, Ö. Turgut, N. Stolt-Ansó, and D. Rueckert. Global and local contrastive learning for joint representations from cardiac mri and ecg. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 217–227, 2025.

[61] H. Sun, S. Lu, H. Wang, Q.-G. Chen, Z. Xu, W. Luo, K. Zhang, and M. Li. Mdp3: A training-free approach for list-wise frame selection in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24090–24101, 2025.

[62] X. Tang, J. Qiu, L. Xie, Y. Tian, J. Jiao, and Q. Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29118–29128, 2025.

[63] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[65] S. Wang, T. Niu, R. Yang, D. Liu, X. He, Z. Wen, C. He, X. Hu, and L. Zhang. Videocompressa: Data-efficient video understanding via joint temporal compression and spatial reconstruction. *arXiv preprint arXiv:2511.18831*, 2025.

[66] X. Wang, J. Liu, Z. Wang, X. Yu, J. Wu, X. Sun, Y. Su, A. Yuille, Z. Liu, and E. Barsoum. Keyvid: Keyframe-aware video diffusion for audio-synchronized visual animation. *arXiv preprint arXiv:2504.09656*, 2025.

[67] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *Advances in Neural Information Processing Systems*, volume 36, pages 7594–7611, 2023.

[68] Y. Wang, Y. Li, X. Zhang, X. Liu, A. Dai, A. B. Chan, and Z. Cui. Edit temporal-consistent videos with image diffusion model. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(12), 2024.

[69] Z. Wang, Z. Yuan, X. Wang, Y. Li, T. Chen, M. Xia, P. Luo, and Y. Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, 2024.

[70] Y. Xia, L. Zhang, N. Ravikumar, R. Attar, S. K. Piechnik, S. Neubauer, S. E. Petersen, and A. F. Frangi. Recovering from missing data in population imaging – cardiac mr image imputation via conditional generative adversarial nets. *Medical Image Analysis*, 67:101812, 2021.

[71] J. Xiao, W. Zheng, W. Wang, Q. Xia, Z. Yan, Q. Guo, X. Wang, S. Nie, and S. Zhang. Slice2mesh: 3d surface reconstruction from sparse slices of images for the left ventricle. *IEEE Transactions on Medical Imaging*, 44(3):1541–1555, 2025.

[72] J. Xing, M. Xia, Y. Zhang, H. Chen, W. Yu, H. Liu, G. Liu, X. Wang, Y. Shan, and T.-T. Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417, 2024.

[73] X. You, R. Yang, C. Zhang, Z. Jiang, J. Yang, and N. 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 (ICCV)*, pages 28010–28020, 2025.

[74] J. Zhang, X. Yang, X. Luo, and L. Shen. Echovdiff: Cardiac-cycle echocardiography video generation from arbitrary single frame. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9040–9050, 2026.

[75] L. Zhang, S. Cai, M. Li, G. Wetzstein, and M. Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[76] L. Zhang, M. Pereañez, C. Bowles, S. Piechnik, S. Neubauer, S. Petersen, and A. Frangi. Missing slice imputation in population cmr imaging via conditional generative adversarial nets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 651–659, 2019.

[77] X. Zhang, J. Liao, S. Zhang, F. Meng, X. Wan, J. Yan, and Y. Cheng. VideoREPA: Learning physics for video generation through relational alignment with foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[78] Z. Zhang, J. Liao, M. Li, Z. Dai, B. Qiu, S. Zhu, L. Qin, and W. Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2063–2073, 2025.

[79] Z. Zhang, F. Long, Z. Qiu, Y. Pan, W. Liu, T. Yao, and T. Mei. Motionpro: A precise motion controller for image-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27957–27967, 2025.

[80] B. Zheng, K. Chen, Y. Yao, Z. Zeng, X. Jiang, H. Wang, J. Lasenby, and X. Jin. Autokeyframe: Autoregressive keyframe generation for human motion synthesis and editing. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH Conference Papers '25, 2025.

[81] X. Zhou, Y. Huang, W. Xue, H. Dou, J. Cheng, H. Zhou, and D. Ni. Heartbeat: Towards controllable echocardiography video synthesis with multimodal conditions-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 361–371, 2024.

[82] X. Zhou, J. Liu, S. Yu, H. Yang, C. Li, T. Tan, and S. Wang. A diffusion-driven temporal super-resolution and spatial consistency enhancement framework for 4d mri imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12, 2025.

[83] Z. Zhu, H. Xu, Y. Luo, Y. Liu, K. Sarkar, Z. Yang, and Y. You. Focus: Efficient keyframe selection for long video understanding. *arXiv preprint arXiv:2510.27280*, 2025.

Appendix

A Datasets

The details of the CMR and ECHO datasets used in our experiments are provided in Table 5.

(i) 3D+t CMR dataset. We use the UK Biobank (UKBB) [52] dataset, which contains 4,379 paired short-axis CMR sequences and ECG recordings, with 50 frames per subject. We use 3,503 subjects for training and 438 each for validation and testing. All CMR volumes are resized to $128 \times 128 \times 13$ and temporally aligned to a single cardiac cycle. For each subject, resting 12-lead ECG signals were parsed from the XML-encoded UK Biobank records and converted to millivolts using acquisition metadata. All leads were synchronously sampled at 500 Hz. R-peaks were detected using NeuroKit2, and a representative R–R interval was selected to define a single-cycle electrical description.

(ii) ECHO datasets. We use two public echocardiography datasets, EchoNet-Dynamic [49] and EchoNet-Pediatric [56], for evaluation. EchoNet-Dynamic contains only A4C views, and we adopt its official train/validation/test split. EchoNet-Pediatric contains both A4C and PSAX views, which we treat as two subsets, i.e., Pediatric (A4C) and Pediatric (PSAX); following its 10-fold protocol, we use folds 1–8 for training, fold 9 for validation, and fold 10 for testing. Each ECHO video is resized to 112×112 with 125 frames. Both datasets provide expert-labeled ejection fraction (EF) scores, together with annotated end-diastolic (ED) and end-systolic (ES) frames.

Table 5: Details of the CMR and ECHO datasets used in our experiments.

Datasets	Modality / View	Train.	Val.	Test.	Raw Average Frame Count
CMR Dataset					
UKBB [52]	3D+t CMR	3503	438	438	50
ECHO Datasets					
EchoNet-Dynamic [49]	A4C	7465	1288	1277	177
EchoNet-Pediatric (A4C) [56]	A4C	2580	336	368	101
EchoNet-Pediatric (PSAX) [56]	PSAX	3559	448	519	103

B More Training Details

B.1 Implementation Details

Tokenizer Training. Adam is used with batch sizes 4 (CMR) and 32 (ECHO). We set σ to 0.2. The weights λ_1 and λ_2 are set to 0.1 and 0.05, respectively. The learning rates are `2e-4/5e-5`. The visual tokenizer, LeanVAE [12], is optimized with a reconstruction objective $\mathcal{L}_{\text{rec}}$ composed of an L_1 reconstruction loss, a perceptual loss, an adversarial loss, and a KL regularization term following the original setting. We use LeanVAE [12] with a compression ratio of $4 \times 8 \times 8$ and 16 latent channels. For CMR, we treat each 3D volume at a cardiac time point as a video by regarding the depth axis as the tokenizer's temporal axis; accordingly, each volume is compressed along the depth, height, and width dimensions with a factor of $4 \times 8 \times 8$, while the cardiac temporal dimension is kept consistent. For ECHO, we follow the standard video-tokenizer pipeline and compress each video sample along the temporal, height, and width dimensions with a $4 \times 8 \times 8$ ratio. For temporal alignment, we map both visual and ECG tokens to the same shared temporal length T. Specifically, for CMR, the ECG input has length 500 and is tokenized by the ECG ViT with patch size 10, yielding 50 temporal tokens, which matches the temporal length of the visual latent sequence produced by the LeanVAE encoder; similarly, for ECHO, the ECG input has length 128 and is tokenized with patch size 4 into 32 temporal tokens, matching the 32-frame visual latent sequence. Our tokenizer training for ECHO is conducted on the combined EchoNet-Dynamic [49] and EchoNet-Pediatric [56] datasets.

Diffusion Model Training. For the diffusion process, the total batch size is 16/64, and training is performed for 200k steps. We optimize the diffusion model using AdamW [42] with a learning rate of `5e-5/1e-4`. All experiments are conducted on 4 NVIDIA A100 GPUs. We precompute and cache the latents and keyframe scores q on disk before diffusion training for efficiency, which significantly reduces training overhead. We further maintain an exponential moving average (EMA) of model

Table 6: Overview of hyperparameter settings for CMR synthesis.

Architecture	
Input dim.	$50 \times 16 \times 16 \times 16 \times 4$
Hidden dim.	768
Num. blocks	12
Num. heads	12
Patch size	2
Params (M)	144.30
FLOPs (G)	639.92
Optimization	
Batch size	$4 \times 4 = 16$
Optimizer	AdamW
Lr	5e−5
(β_1, β_2)	$(0.9, 0.95)$
Lr schedule	Warmup cosine
Training steps	200k
EMA decay	0.9999
Interpolants	
α_t	$1 - t$
σ_t	t
ω_t	σ_t
Training objective	v-prediction
Sampler	Euler-Maruyama
Sampling steps	250
CardioFAD	
Num. keyframes	10
$p(A), p(B)$	$p(A) = 0.5, p(B) = 0.5$
λ_A, λ_B	$\lambda_A = 1.5, \lambda_B = 0.5$

Table 7: Overview of hyperparameter settings for ECHO synthesis.

Architecture	
Input dim.	$32 \times 16 \times 14 \times 14$
Hidden dim.	768
Num. blocks	12
Num. heads	12
Patch size	2
Params (M)	144.06
FLOPs (G)	157.37
Optimization	
Batch size	$16 \times 4 = 64$
Optimizer	AdamW
Lr	1e−4
(β_1, β_2)	$(0.9, 0.95)$
Lr schedule	Warmup cosine
Training steps	200k
EMA decay	0.9999
Interpolants	
α_t	$1 - t$
σ_t	t
ω_t	σ_t
Training objective	v-prediction
Sampler	Euler-Maruyama
Sampling steps	250
CardioFAD	
Num. keyframes	8
$p(A), p(B)$	$p(A) = 0.5, p(B) = 0.5$
λ_A, λ_B	$\lambda_A = 1.5, \lambda_B = 0.5$

weights with a decay rate of 0.9999. During sampling, we use the Euler–Maruyama sampler with 250 steps. During inference, keyframes are selected deterministically as the hard top-K indices of the learned importance scores q, rather than being stochastically sampled. The detailed hyperparameter setups of our diffusion model are provided in Tables 6 and 7.

Observation Setup. In all sparse-observation settings, the observation is randomly sampled. For **Volume2Seq**, we observe one 3D volume at a random time index. For **Slice2Seq**, we observe one 2D slice at a random depth index. For **Frame2Video**, we observe one frame at a random time index.

B.2 Details of Two-stage Generation

Video DiT Architecture. We adopt a factorized Video DiT with alternating spatial and temporal transformer blocks to separately model subject-specific anatomy and cardiac dynamics. The architecture processes both 3D+t CMR and 2D+t ECHO latents using a unified conditioning strategy:

- *Input Formulation:* The observed reference latent (volume, slice, or frame) is replicated along the temporal axis and concatenated with the noisy latent along the channel dimension before patch embedding. We apply fixed spatial/temporal positional embeddings alongside modality-specific rotary positional encodings (3D for CMR, 2D for ECHO).
- *Spatial Blocks (Anatomy Preservation):* To maintain anatomical consistency, the reference latent is mean-pooled and injected as a global condition via AdaLN. For the Slice2Seq task, the specific slice depth index is concatenated to this spatial condition to provide spatial awareness.
- *Temporal Blocks (Motion Generation):* Temporal dynamics are guided by three signals: the observation time index, the keyframe probability q, and the physiological time-series embedding (ECG for CMR, pseudo-ECG for ECHO). Their pooled representation is injected globally via AdaLN, while the frame-wise time-series embeddings are additionally injected via cross-attention to provide fine-grained, temporally varying cardiac guidance.

Unified Masked Diffusion. We train a single Video DiT using a unified masked diffusion objective for both keyframe generation and non-keyframe interpolation, rather than training separate models.

Given a binary keyframe mask $m^{kf} \in \{0,1\}^T$, we randomly sample one of two generation modes at each training iteration:

- *Mode A (Keyframe Generation):* Supervised using the mask $m = m^{kf}$.
- *Mode B (Frame Interpolation):* Supervised using the complementary mask $m = 1 - m^{kf}$.

Following the transport-based diffusion formulation, the network predicts the target velocity field conditioned on all available signals (i.e., the observed reference latent, observation index, time-series embeddings, and keyframe probability). Crucially, the temporal mask m is applied exclusively to the loss function, allowing the same denoiser to supervise different subsets of frames:

$$\mathcal{L}_{diff} = \lambda_{mode}\text{Mean}[m \odot (v_\theta - u_\tau)^2] \tag{8}$$

In practice, both modes are sampled with equal probability (0.5). To encourage the model to prioritize accurate anchor generation before refining intermediate dynamics, we assign a larger loss weight to Mode A ($\lambda_A = 1.5$) than to Mode B ($\lambda_B = 0.5$). This complementary masking strategy enables a single network to efficiently and jointly learn both stages.

B.3 Pseudo-ECG Construction and Usage for ECHO

For ECHO datasets lacking native ECG recordings, we construct a surrogate 1D physiological prior (pseudo-ECG) to provide coarse, normalized cardiac-cycle conditioning.

Mathematical Formulation. We standardize the cardiac cycle to length N (we set $N = 128$) with a normalized phase $\phi \in [0, 1)$. Setting the ED phase to $\phi_{ED} = 0$ and the ES phase to $\phi_{ES} = 0.3$ based on standard physiological proportions, we model ventricular volume changes using a smooth cosine interpolation $g(\phi)$:

$$g(\phi) = \begin{cases} \frac{1}{2}[1 + \cos(\pi \frac{\phi}{\phi_{ES}})], & 0 \le \phi \le \phi_{ES} \\ \frac{1}{2}[1 - \cos(\pi \frac{\phi - \phi_{ES}}{1 - \phi_{ES}})], & \phi_{ES} < \phi < 1 \end{cases} \tag{9}$$

After min-max normalization, the patient-specific anatomical areas mathematically cancel out, rendering the normalized amplitude exactly equivalent to the canonical template $g(\phi)$. The final 1D pseudo-ECG token at step n is simply concatenated as $e_n = [\phi_n, g(\phi_n), \Delta g(\phi_n)]$.

Patient-Agnostic and Zero-Leakage. Crucially, because e_n reduces to a universal, patient-agnostic template, it requires absolutely no patient-specific annotations (e.g., actual ED/ES frame indices or contour areas) during inference. In our extreme one-shot clinical setting, the model conditions exclusively on this generic periodic signal, strictly ensuring zero temporal information leakage at test time. Moreover, Table 17 and Appendix C.9 (*Robustness to ECG Perturbations*) provide controlled ablations with degraded/perturbed ECG guidance, showing that the performance gains do not hinge on the exact pseudo-ECG quality and that the proposed AAP-driven alignment remains effective beyond this weak auxiliary cue. A controlled attribution study in Appendix C.6 further disentangles this temporal guidance, showing that the ECHO gains arise from valid generic phase conditioning and two-stage keyframe anchoring rather than patient-specific ECG information.

B.4 Details of EF Estimation

For cardiac functional preservation evaluation, we follow the standard echocardiography setting and treat EF estimation as a downstream regression task that predicts a continuous ejection fraction (EF) value from an input video [57, 58]. Specifically, the EF dataset consists of video–label pairs (V, y), where V denotes an echocardiographic sequence and y is the corresponding ground-truth EF score. A regression model $\text{Reg}(\cdot)$ takes the full sequence as input and outputs an estimate $\hat{y} = \text{Reg}(V)$, and is trained by minimizing the mean-squared error, $\mathcal{L}_{\text{reg}} = (y - \hat{y})^2$. We apply the same protocol to evaluate generated videos, using downstream EF performance as a proxy for whether the synthesized dynamics preserve clinically relevant functional cues. Better EF estimation performance therefore suggests stronger cardiac functional consistency in the generated sequences.

Table 8: **Frozen visual tokenizer validation.** The visual tokenizer is trained independently and frozen before learning the ECG encoder, selector, alignment module, and diffusion model.

Methods	Latent MSE↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
Uniform keyframes	0.023	25.94	0.572	0.081	40.41	188.36
Random keyframes	0.027	25.17	0.554	0.091	45.59	251.03
GT ED/ES	0.020	26.20	0.572	0.083	42.76	202.41
CardioFAD (One-stage)	–	25.14	0.556	0.089	46.82	232.17
CardioFAD (Frozen Tokenizer)	0.018	27.56	0.592	0.070	31.84	162.93
CardioFAD (Joint Tokenizer)	**0.014**	**27.73**	**0.596**	**0.067**	**26.92**	**137.70**

Table 9: **Cardiac functional consistency** evaluated by EF estimation for 2D+t ECHO synthesis on EchoNet-Pediatric and its A4C/PSAX subsets [56].

Methods	**EchoNet-Pediatric (A4C)**			**EchoNet-Pediatric (PSAX)**		
	R^2↑	MAE↓	RMSE↓	R^2↑	MAE↓	RMSE↓
General Video Generation						
VideoComposer [67]	0.68	4.93	6.73	0.62	4.88	7.21
Tora [78]	0.66	4.89	6.86	0.69	4.59	6.52
ECHO Video Generation						
HeartBeat [81]	0.71	4.62	6.29	0.65	4.40	6.94
EchoNet-Synthetic [58]	0.70	4.57	6.44	0.68	4.19	6.60
ECHOPulse [34]	0.73	4.39	6.07	0.72	4.06	6.15
CardioFAD (One-stage)	0.73	4.35	6.02	0.73	4.10	6.02
CardioFAD (Ours)	**0.75**	**4.23**	**5.86**	**0.76**	**3.98**	**5.79**

C Additional Experimental Results

C.1 Frozen Visual Tokenizer Validation

One potential concern is that the learned keyframes may be partially coupled with the jointly optimized visual latent space. To investigate whether our selected keyframes remain informative under an independently fixed representation, we conduct an additional validation with a frozen visual tokenizer on the EchoNet-Dynamic dataset [49] in Table 8. Concretely, we first train our visual encoder, LeanVAE tokenizer [12], using only the standard reconstruction objective for the same number of training steps as the full model, without the TSP, AAP, or keyframe selector. We then freeze the visual encoder and decoder, and train the ECG encoder, keyframe selector and alignment module on top of this fixed latent space. In this setting, the selector cannot reshape the visual representation to satisfy the TSP objective, and must identify keyframes that are sufficient under an independently learned latent trajectory.

It can be observed from Table 8 that CardioFAD with a frozen tokenizer still outperforms random, uniform, and GT ED/ES keyframes in both latent trajectory reconstruction and synthesis quality. It also clearly improves over the one-stage baseline, indicating that keyframe anchoring remains effective even when the visual latent space is fixed. The joint-tokenizer model further improves performance, suggesting that joint optimization enhances representation compatibility, while the frozen-tokenizer validation demonstrates that the learned keyframes are not merely artifacts of co-adapted latent representations.

C.2 Additional Cardiac Functional Evaluation

Functional consistency across pediatric ECHO views. We further report cardiac functional consistency evaluation for 2D+t ECHO synthesis on EchoNet-Pediatric and its A4C/PSAX subsets in Table 9. Across both views, our full model achieves the strongest overall downstream EF estimation performance, while the one-stage variant remains competitive, confirming the benefit of the proposed Focus–Align–Diffuse design. The improvements over both general video generation and ECHO-

Table 10: **EF-stratified functional preservation** on EchoNet-Dynamic.

Metric	Reduced $\text{EF} \leq 40$	Mildly reduced $40 < \text{EF} < 50$	Preserved-range $\text{EF} \geq 50$	All
n	161	125	991	1,277
Real-video MAE ↓	4.81	4.49	3.76	3.96
One-stage MAE ↓	5.84	5.29	4.80	4.98
CardioFAD MAE ↓	**5.10**	**4.65**	**3.96**	**4.17**
Bias vs. GT	0.72	0.58	0.15	0.26
Paired Gen–Real MAE ↓	2.13	1.79	1.40	1.53

Table 11: **Direct motion preservation** on EchoNet-Dynamic.

Metric	All stratified cases	Reduced EF	High motion heterogeneity
n	60	20	15
Area-curve MAE (One-stage) ↓	0.063	0.078	0.095
Area-curve MAE (CardioFAD) ↓	**0.048**	**0.051**	**0.067**
Segmental-motion MAE (One-stage) ↓	0.119	0.140	0.182
Segmental-motion MAE (CardioFAD) ↓	**0.097**	**0.116**	**0.141**

specific baselines indicate that CardioFAD better preserves EF-related functional information in the synthesized sequences.

Pathology-specific motion patterns. We further examine whether CardioFAD preserves EF-related functional differences and regional motion patterns on EchoNet-Dynamic. We first divide all 1,277 test subjects into reduced ($\text{EF} \leq 40$), mildly reduced ($40 < \text{EF} < 50$), and preserved-range ($\text{EF} \geq 50$) groups. Using the same frozen EF regressor, Table 10 reports the real-video MAE, generated-video MAE, signed bias, and paired generated–real MAE. CardioFAD improves over the one-stage model in every EF group, while maintaining only modest bias, supporting the preservation of EF-related severity differences.

We further evaluate direct motion consistency on a stratified subset of 60 subjects (20 per EF group) using a frozen LV segmenter. The 15 cases with the highest real-video segmental-motion heterogeneity are additionally evaluated as a challenging subset. Area-curve MAE measures the discrepancy between ED-normalized LV area trajectories over the cardiac cycle, while segmental-motion MAE compares six A4C endocardial-segment shortening profiles from ED to ES. CardioFAD reduces both errors across all evaluated subsets in Table 11, including reduced-EF and high-heterogeneity cases. The consistent gains suggest improved preservation of global contraction–relaxation and heterogeneous regional motion.

C.3 Multi-seed Robustness Analysis

To assess robustness to training randomness, we train and evaluate all methods using three independent random seeds on EchoNet-Dynamic. As reported in Table 12, CardioFAD consistently outperforms ECHOPulse [34] and the one-stage variant across all metrics. In particular, its 2.08 dB PSNR gain over ECHOPulse is substantially larger than the across-seed variation, with a two-sided Welch's t-test yielding $p < 0.001$. These results confirm the robustness of our improvements and the benefit of the two-stage design.

C.4 Additional Ablation Studies

Principle-guided Keyframe Learning for CMR Synthesis. We conduct an additional ablation on principle-guided keyframe learning for CMR Volume2Seq synthesis on the UKBB dataset (Table 13). Removing either the TSP or the AAP consistently degrades performance across all metrics, and disabling both leads to the most pronounced deterioration, indicating that the two principles are complementary. We further compare different keyframe selection strategies and find that uniform or random keyframes perform notably worse than our learned keyframes. Overall, these results verify

Table 12: **Multi-seed robustness analysis on EchoNet-Dynamic.** Results are reported as mean ± standard deviation over three independent random seeds.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
ECHOPulse [34]	$25.60_{\pm 0.19}$	$0.567_{\pm 0.003}$	$0.081_{\pm 0.004}$	$37.24_{\pm 2.5}$	$194.19_{\pm 11.5}$
CardioFAD (One-stage)	$25.17_{\pm 0.16}$	$0.558_{\pm 0.003}$	$0.090_{\pm 0.003}$	$46.52_{\pm 2.9}$	$235.08_{\pm 9.8}$
CardioFAD (Ours)	$\mathbf{27.68}_{\pm 0.20}$	$\mathbf{0.595}_{\pm 0.004}$	$\mathbf{0.068}_{\pm 0.004}$	$\mathbf{26.48}_{\pm 2.2}$	$\mathbf{139.54}_{\pm 12.8}$

Table 13: **Principle-guided keyframe learning ablation** for CMR Volume2Seq synthesis.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
Keyframe Learning Principle					
w/o TSP	29.07	0.836	0.098	104.53	1226.37
w/o AAP	30.53	0.861	0.082	81.63	910.04
w/o TSP & w/o AAP	28.21	0.827	0.111	134.54	1507.13
Keyframe Selection Strategy					
w/ uniform keyframes	29.86	0.852	0.094	97.41	1028.59
w/ random keyframes	28.63	0.834	0.105	121.86	1379.10
CardioFAD	**32.41**	**0.885**	**0.064**	**59.96**	**736.01**

that principled keyframe learning is crucial for achieving high-fidelity and temporally consistent synthesis under extreme sparsity.

Effect of Number of Keyframes K. We further analyze how the number of learned keyframes affects CardioFAD in Tables 14 and 15. As shown, a moderate number of keyframes provides the best trade-off between trajectory coverage and focused anchor selection. On CMR Volume2Seq synthesis, performance steadily improves as K increases from 2 to 10, with K=10 achieving the best results across all metrics. However, further increasing K to 15 degrades both fidelity and temporal consistency, indicating that too many anchors may introduce redundancy and dilute the focus on motion-salient phases. A similar trend is observed on EchoNet-Dynamic [49], where K=8 yields the best overall performance, while both smaller settings (K=2, 4) and a larger setting (K=12) lead to inferior results. These results support our design choice of using compact keyframe sets, with K=10 for CMR and K=8 for ECHO, to capture sufficient cardiac dynamics without weakening the anchor-based generation process.

Effect of Conditioning Guidance. We further ablate the conditioning signals used in our video DiT in Tables 16 and 17. It can be observed that each condition contributes to the generation quality, and removing any component leads to degradation across both CMR and ECHO settings. Specifically, removing ECG conditioning causes the most severe drop, especially in FID and FVD, indicating that physiological time-series guidance is essential for resolving temporal ambiguity under sparse visual observations. Removing cross-modal alignment also degrades performance, verifying that ECG signals cannot be fully exploited without explicitly aligning them with the visual cardiac trajectory. In addition, removing keyframe probabilities weakens the model's awareness of motion-salient anchors, while removing the reference-frame index makes it harder to localize the sparse observation within the cardiac cycle. To sum up, these ablation results confirm that CardioFAD benefits from jointly conditioning the diffusion model on physiological signals, aligned cross-modal representations, keyframe saliency, and observation position, rather than relying on any single source of guidance.

C.5 Qualitative Results of Ablation Studies

Figs. 11 and 12 compare our CardioFAD with the one-stage variant and uniform keyframe sampling under the same sparse observation. Across both CMR and ECHO, CardioFAD yields sharper anatomy and more coherent temporal evolution, especially for frames between keyframes and the hard non-keyframes where motion is most ambiguous. In contrast, uniform keyframes and the one-stage variant exhibit stronger blurring and localized artifacts, which is reflected by larger residuals in the difference maps.

Table 14: **Ablating the number of keyframes** for CMR Volume2Seq synthesis.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
$K = 2$	30.29	0.863	0.083	86.26	1063.45
$K = 5$	30.90	0.867	0.079	80.73	945.92
$K = 8$	32.27	0.882	0.068	65.93	825.14
K = 10	**32.41**	**0.885**	**0.064**	**59.96**	**736.01**
$K = 15$	31.16	0.873	0.073	84.67	907.22

Table 15: **Ablating the number of keyframes** on EchoNet-Dynamic.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
$K = 2$	26.18	0.574	0.077	36.30	173.79
$K = 4$	27.11	0.583	0.072	30.64	148.27
K = 8	**27.73**	**0.596**	**0.067**	**26.92**	**137.70**
$K = 12$	27.36	0.583	0.069	28.24	161.08

Table 16: **Condition ablation** for CMR Volume2Seq synthesis.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
w/o alignment	30.89	0.864	0.081	77.85	913.72
w/o ECG condition	29.24	0.834	0.099	107.39	1281.46
w/o keyframe prob.	31.12	0.869	0.074	80.92	866.04
w/o reference index	31.37	0.873	0.072	77.05	901.53
CardioFAD	**32.41**	**0.885**	**0.064**	**59.96**	**736.01**

Table 17: **Condition ablation** on EchoNet-Dynamic.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
w/o alignment	25.71	0.562	0.080	39.54	209.67
w/o ECG condition	25.26	0.559	0.088	45.98	243.62
w/o keyframe prob.	26.82	0.584	0.073	32.28	169.74
w/o reference index	26.27	0.576	0.079	40.46	188.75
CardioFAD	**27.73**	**0.596**	**0.067**	**26.92**	**137.70**

C.6 Attribution of Temporal Guidance

As reported in Table 18, we conduct a controlled attribution study by varying only the temporal condition under otherwise matched settings. On UKBB [52], paired real ECG consistently outperforms subject-shuffled, canonical, and null ECG, indicating that generic temporal structure is beneficial while correct patient-specific correspondence provides additional gains. On EchoNet-Dynamic [49], the correct phase template outperforms both random-offset and null templates, while the two-stage model further improves over its matched one-stage counterpart. These results indicate that patient-specific ECG provides additional value on UKBB, whereas the ECHO gains arise from valid generic phase conditioning together with two-stage keyframe anchoring.

C.7 Conditional Fidelity–Diversity Analysis

One-shot cardiac synthesis is inherently one-to-many, and our notion of *recoverability* refers to trajectory reconstruction sufficiency rather than unique identification of the underlying motion.

Experimental Setup. To examine the resulting conditional uncertainty, we evaluate 100 EchoNet-Dynamic [49] test cases and generate eight samples for each fixed sparse observation. We keep the observed frame, observation index, pseudo-ECG, and hard top-K keyframes fixed, while varying only the rectified-flow sampling seed.

Evaluation Metrics. We report Avg. LPIPS for average sample fidelity, Best-of-8 LPIPS for coverage of the held-out realization, and Pairwise LPIPS and normalized motion diversity for appearance and motion variation across samples. Uncertainty–error ρ denotes the Spearman correlation between frame-wise sample disagreement and reconstruction error, while EF MAE and within-case EF standard deviation evaluate functional fidelity and stability.

As reported in Table 19, CardioFAD improves both average and best-of-eight fidelity while reducing unconstrained appearance and motion variation. Its uncertainty–error correlation increases from 0.29 to 0.50, while EF MAE and within-case EF variability decrease to 4.21 and 1.35, respectively, indicating that the remaining stochastic variation is better localized to difficult frames and functionally more stable. Overall, CardioFAD prioritizes faithful and phase-consistent reconstruction over maximal diversity rather than attempting to model the complete conditional posterior.

C.8 Sensitivity to the Observed-Frame Phase

To examine sensitivity to the location of the sparse observation within the cardiac cycle, we conduct a phase-stratified evaluation on EchoNet-Dynamic [49]. All videos are temporally resampled to 125 frames, and we use the same trained checkpoint and inference protocol while changing only the observed frame and its temporal index. ED/ES annotations are used only to define the phase groups

Table 18: **Attribution of temporal guidance on UKBB and EchoNet-Dynamic.** We vary only the temporal condition under otherwise matched settings to disentangle patient-specific ECG information, generic phase guidance, and two-stage keyframe anchoring.

Dataset	Temporal condition	Patient-specific	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
UKBB	**Paired real ECG**	✓	**32.41**	**0.885**	**0.064**	**59.96**	**736.01**
	Subject-shuffled real ECG	×	31.12	0.870	0.076	91.48	958.23
	Canonical training ECG	×	30.74	0.862	0.083	98.77	1046.28
	Null ECG	×	29.24	0.834	0.099	107.39	1281.46
EchoNet-Dynamic	**Correct phase template**	×	**27.73**	**0.596**	**0.067**	**26.92**	**137.70**
	Random-offset template	×	25.59	0.563	0.085	37.93	217.08
	Null template	×	25.26	0.559	0.088	45.98	243.62
	Correct template, one-stage	×	25.14	0.556	0.089	46.82	232.17

Table 19: **Conditional fidelity–diversity analysis on EchoNet-Dynamic.** Eight samples are generated for each fixed sparse observation by varying only the rectified-flow sampling seed.

Methods	Avg. LPIPS↓	Best-of-8 LPIPS↓	Pairwise LPIPS	Norm. motion diversity	Uncertainty–error↑	EF MAE↓	Within-case EF std.↓
CardioFAD (One-stage)	0.091	0.085	0.035	0.40	0.29	5.03	2.21
CardioFAD (Ours)	**0.070**	**0.064**	**0.025**	**0.27**	**0.50**	**4.21**	**1.35**

for this analysis and are never provided to the model. Using circular temporal indexing, we define four observation settings:

- **Near ED**: frames within ± 2 frames of the annotated ED frame, forming a five-frame window.
- **Near ES**: frames within ± 2 frames of the annotated ES frame, also forming a five-frame window.
- **Contraction**: intermediate frames between the near-ED and near-ES windows along the ED-to-ES trajectory.
- **Relaxation**: intermediate frames between the near-ES window and the next near-ED window along the ES-to-ED trajectory.

Table 20 indicates that near-ED and near-ES observations perform slightly better than intermediate contraction and relaxation phases, consistent with motion-extremal states providing stronger trajectory constraints. Nevertheless, the maximum variation is limited to only 0.24 dB in PSNR, 15.9 in FVD, and 0.18 in EF MAE. These results indicate that CardioFAD is only mildly sensitive to the observed-frame phase, with phase conditioning and two-stage anchoring largely compensating for less informative observation positions.

C.9 Robustness to ECG Perturbations

ECG Perturbations for CMR Synthesis. In real-world clinical settings, auxiliary ECG signals can be corrupted by measurement noise, missing leads, or temporal desynchronization. To assess the intrinsic robustness of our method, we evaluate models trained only on clean data under controlled ECG perturbations only at inference time on UKBB.

(i) Signal Noise. We inject Gaussian noise into the ECG condition with probability $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ (with a fixed noise magnitude σ, or equivalently matched to a target SNR), simulating stochastic acquisition artifacts. Concretely, for each time index t, we apply $\tilde{\mathbf{e}}_t = \mathbf{e}_t + m_t \cdot \sigma \boldsymbol{\epsilon}_t$, where $m_t \sim \text{Bernoulli}(p)$ and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

(ii) Missing Leads. To mimic sensor detachment, we randomly drop $L \in \{1, 2, 4, 6, 8\}$ leads by zero-masking the corresponding channels, while keeping the input dimensionality unchanged.

(iii) Phase Jitter. To emulate hardware synchronization lag, we temporally shift the ECG sequence by $\delta \in \{1, 2, 4, 6, 8\}$ frames (using circular shift or padding-based shift), introducing controlled misalignment between ECG and image dynamics.

As shown in Fig. 9, while all methods degrade under ECG perturbations, CardioFAD degrades most gracefully. The performance gap is especially pronounced under missing leads and phase jitter. This resilience indicates that our AAP acts as a robust soft constraint; rather than overfitting to exact

Table 20: **Sensitivity to the observed-frame phase** on EchoNet-Dynamic under matched inference settings.

Observed-frame phase	PSNR↑	FVD↓	EF MAE↓
Near ED	**27.92**	**128.92**	4.10
Near ES	27.84	135.04	**4.08**
Contraction	27.72	140.57	4.23
Relaxation	27.68	144.82	4.26
Random observation	27.73	137.70	4.17

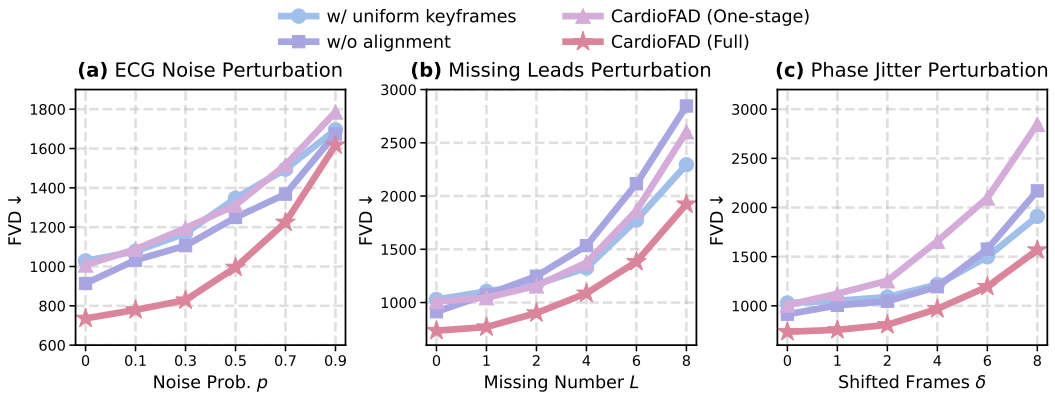


Figure 9: **Robustness to ECG perturbations on UKBB (CMR).** FVD (lower is better) under three inference-time ECG corruptions: Gaussian noise injection with probability p, random lead dropout with L missing leads, and temporal phase jitter with δ shifted frames. CardioFAD (Full) degrades most gracefully and remains consistently best across perturbation severities.

channel-wise amplitudes or high-frequency fluctuations, CardioFAD successfully anchors generation using macro-level rhythm and phase cues, supporting robustness to the evaluated conditioning corruptions.

Pseudo-ECG Perturbations for ECHO Synthesis. For ECHO, ECG is unavailable and we condition the model on a lightweight pseudo-ECG token sequence. To evaluate robustness to imperfect surrogate signals, we perturb the pseudo-ECG only at inference time on EchoNet-Dynamic, while keeping all models trained on clean pseudo-ECG unchanged.

(i) Phase Perturbation. We apply a global phase shift $\Delta \in \{0.02, 0.05, 0.10, 0.15\}$ to the phase token, simulating synchronization errors and phase ambiguity in template-based conditioning.

(ii) Component Dropout. We ablate the pseudo-ECG components by dropping one channel at a time (phase / amplitude / derivative) via zero-masking, testing which cues the model relies on most.

(iii) Segment Masking. We randomly mask a portion of pseudo-ECG tokens with mask ratio $r \in \{0.1, 0.3, 0.5, 0.7\}$ to mimic missing or incomplete conditioning sequences.

Similarly, for pseudo-ECG perturbations (Fig. 10), CardioFAD consistently exhibits the slowest degradation. While phase alterations severely impact baselines, our model remains substantially more stable. Furthermore, the relatively mild degradation caused by dropping amplitude/derivative channels or masking segments confirms that CardioFAD primarily leverages coarse phase cues rather than fine-grained waveform details, highlighting its robustness under imperfect surrogate conditions.

C.10 Efficiency Analysis

To validate the practical viability of our proposed CardioFAD, we analyze its computational footprint across both inference and tokenizer training, demonstrating that our principle-guided framework achieves substantial generative quality improvements with negligible memory overhead and acceptable runtime costs.

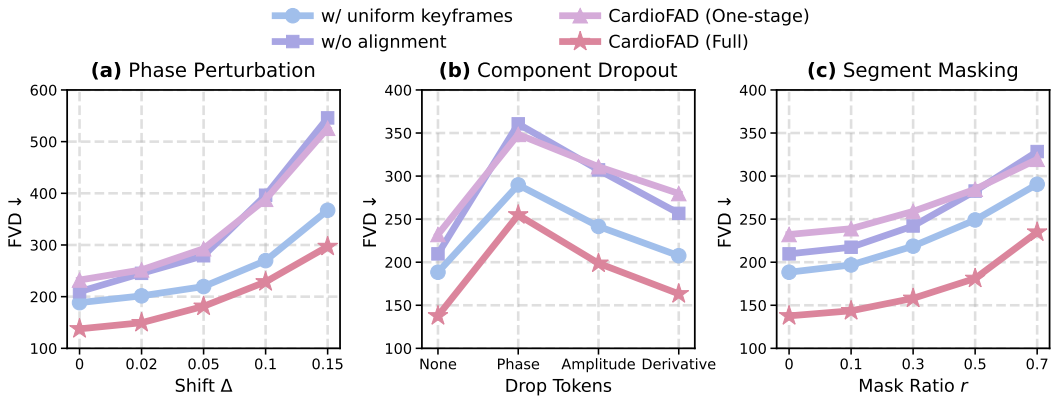


Figure 10: **Robustness to pseudo-ECG perturbations on EchoNet-Dynamic.** FVD (lower is better) under inference-time perturbations of the pseudo-ECG: global phase shifts Δ, component dropout (zero-masking phase/amplitude/derivative), and random segment masking with ratio r. CardioFAD (Full) remains consistently best and degrades most gracefully across perturbation severities.

Table 21: **Inference cost comparison between the one-stage baseline and CardioFAD on CMR and ECHO.** We report end-to-end latency per sample and peak VRAM under standardized settings (batch size $= 1$, single A100 GPU, same sampling steps, and mixed precision).

Methods	CMR (UKBB)		ECHO (EchoNet-Dynamic)	
	Latency (s)↓	Peak VRAM (GB)↓	Latency (s)↓	Peak VRAM (GB)↓
CardioFAD (One-stage)	6.19	14.82	3.34	4.91
CardioFAD (Ours)	12.27	14.84	6.62	4.93

Inference Cost. In terms of inference efficiency (Table 21), CardioFAD introduces a moderate runtime overhead compared with the one-stage baseline, while maintaining nearly identical memory usage. On both CMR and ECHO, the end-to-end latency is approximately doubled, which is consistent with our two-stage sampling design that sequentially performs keyframe generation and sequence completion. Given the clear performance gains of CardioFAD over the one-stage baseline, this increased runtime with virtually no additional VRAM footprint reflects a reasonable quality–efficiency trade-off. Notably, to reduce preprocessing overhead and better reflect the actual generation cost, all latent representations are pre-encoded and cached locally before timing.

Wall-clock Cost of Tokenizer Training. As shown in Table 22, across both CMR and ECHO, incorporating TSP introduces only a minor computational overhead, as the additional parameters/FLOPs mainly come from a lightweight UNet used to reconstruct the full cardiac latent, leading to only marginal increases in per-epoch time and total GPU hours. In contrast, AAP does not add any learnable components and thus keeps the model size and FLOPs essentially unchanged. Despite these minimal training-cost changes, the full model achieves consistently better synthesis quality, yielding higher PSNR and substantially lower FVD compared with the corresponding ablations. Overall, these results indicate a favorable quality–efficiency trade-off, where notable quality gains are obtained with negligible additional wall-clock training burden.

Table 22: **Tokenizer training efficiency vs. downstream synthesis quality.** We compare principle ablations in tokenizer training against the full model. Synthesis quality is reported as the average across the two CMR evaluation settings and across the three ECHO datasets, respectively.

Data	Methods	Tokenizer Training Cost				Synthesis Quality	
		Params (M) ↓	FLOPs (G) ↓	Time/Epoch (s) ↓	Total (GPU h) ↓	PSNR↑	FVD↓
CMR	w/o TSP	48.01	2008.23	480.7	535.1	25.76	1982.53
	w/o AAP	48.56	2011.48	483.0	536.7	27.25	1627.98
	w/o TSP & AAP	48.01	2008.23	480.6	534.0	25.31	2376.82
	CardioFAD (Full)	48.56	2011.48	483.2	536.9	**29.00**	**1291.87**
ECHO	w/o TSP	44.45	247.65	202.3	224.8	27.13	368.04
	w/o AAP	45.00	249.19	203.9	226.6	27.87	331.53
	w/o TSP & AAP	44.45	247.65	201.6	224.0	26.56	428.77
	CardioFAD (Full)	45.00	249.19	204.0	226.7	**28.93**	**239.28**

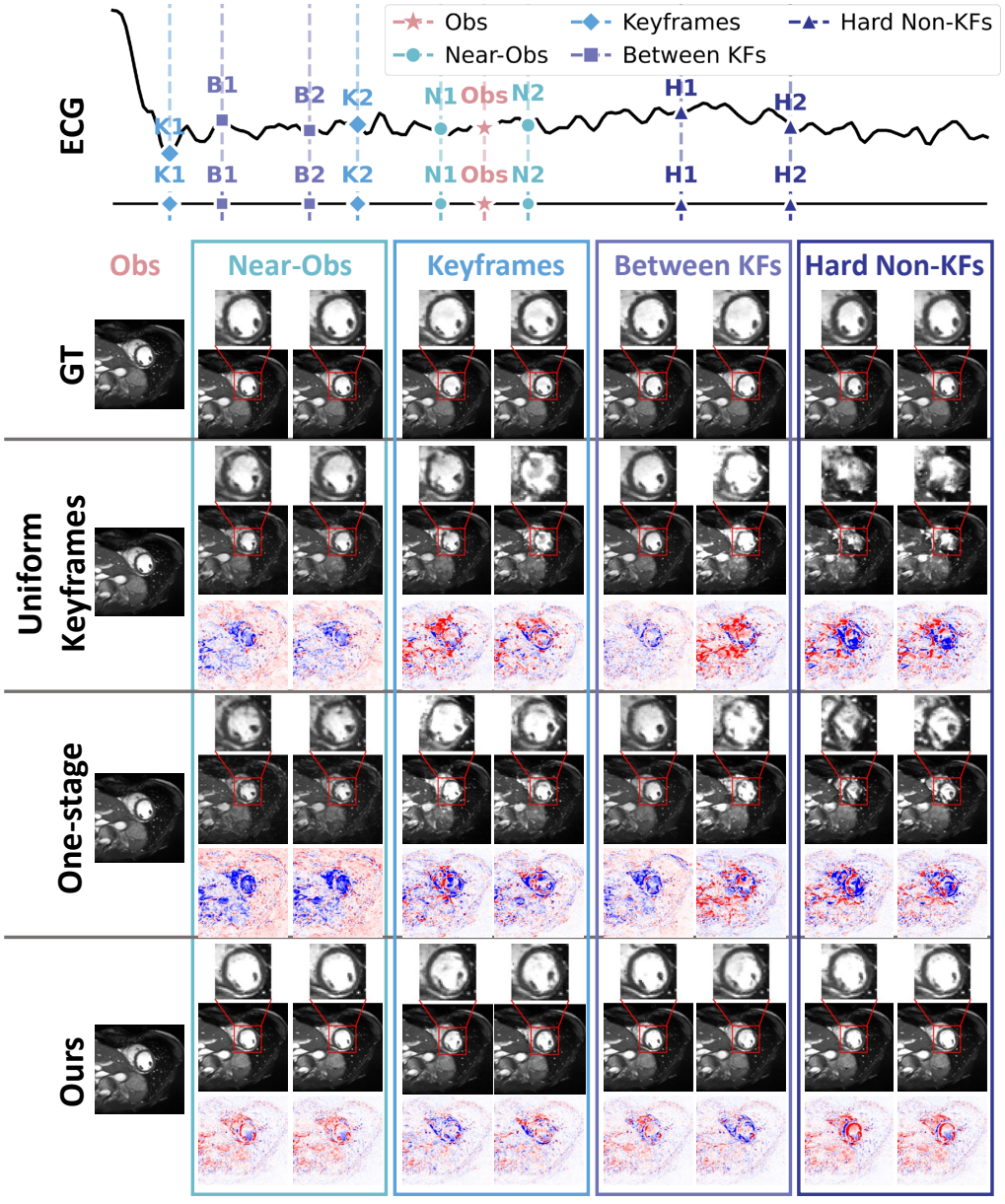


Figure 11: **Qualitative comparisons with variants on CMR.**

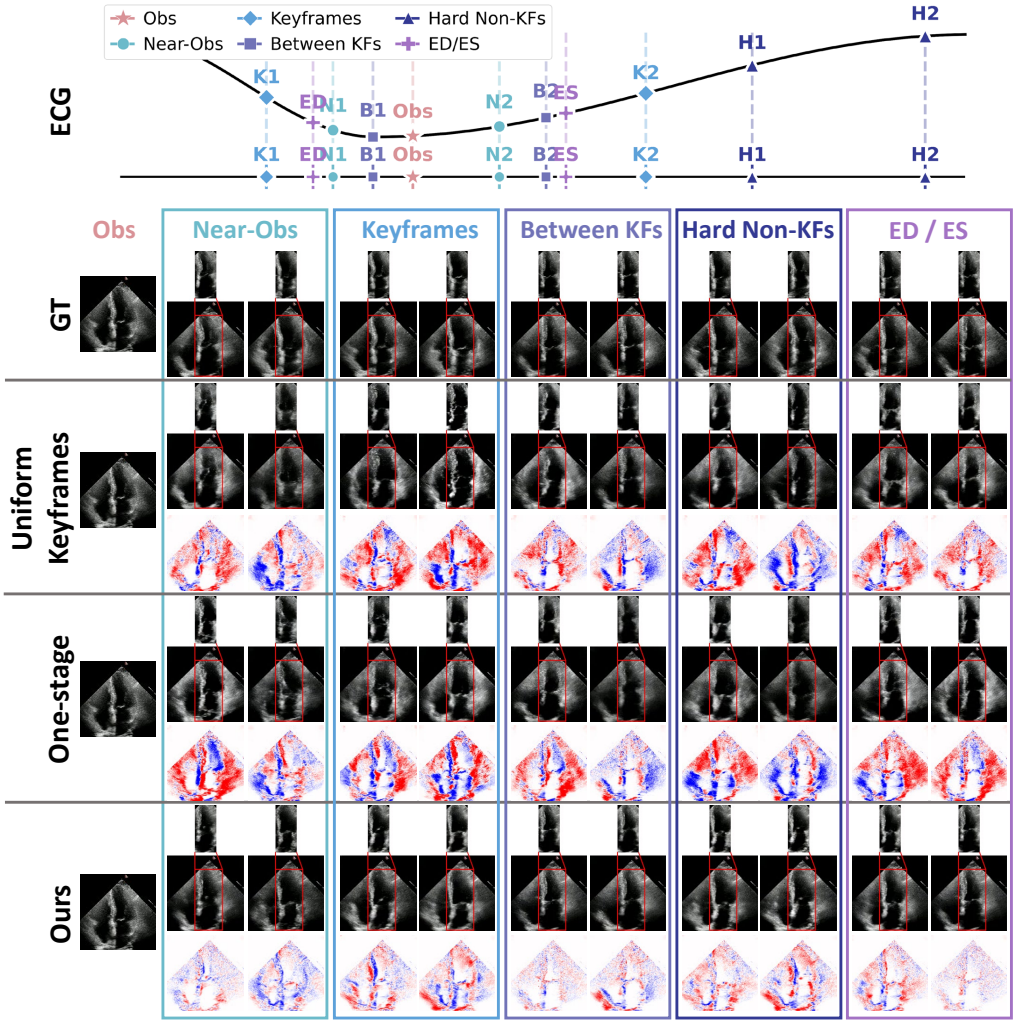


Figure 12: **Qualitative comparisons with variants on ECHO.**

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 main claims presented in our abstract and introduction accurately reflect the contribution of our paper.

   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: We discuss the limitations of our work in Sec. 5.

   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: This paper focuses on proposing a keyframe-to-sequence diffusion framework for one-shot cardiac dynamic synthesis, which does not contain theoretical results.

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: We provide comprehensive experimental and implementation details, as well as hyperparameter settings in Appendix B.1.

   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: [No]

Justification: Some datasets used in this study are not publicly available.

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: We have included the implementation details and the hyperparameters in Sec. 4.1 and Appendix B.1.

   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: We report mean and standard deviation over three independent runs in Appendix C.3.

   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: We have provided the computing cost in Appendix C.10.

   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: We have read and understood the NeurIPS Code of Ethics and have done our best to conform.

   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: [N/A]

    Justification: There is no potential societal impact of our work.

    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: [Yes]

    Justification: Our experiments do not involve this aspect.

    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: We cite the original paper that produced the code package or dataset.

    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: [N/A]

Justification: The paper does not release new assets.

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: This paper does not involve this aspect.

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: [N/A]

Justification: This paper does not involve this aspect.

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: This work does not incorporate any LLMs as part of the core methodology. LLMs were used only for writing and editing suggestions in this paper.

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.