From Heartbeat to Cardiochoreography: A Mechanics Foundation Model for Individualized 4D Cardiac Motion Generation Conditioned on Electrophysiology

Ziquan Wei **Tingting Dan** **Guorong Wu***
Departments of Computer Science and Psychiatry
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
ziquanw@cs.unc.edu;{Tingting_Dan,grwu}@med.unc.edu

Abstract

The electrocardiogram (ECG) is a time series that records the electrical activity driving cardiac motion through electromechanical coupling (EMC), where the depolarization captured by ECG triggers each heartbeat. The temporal delay and low spatial resolution of such multichannel dynamic data often cause temporal and spatial misalignments along the physical motion, limiting the application of ECG data to event/segment-based downstream clinic practice. Although many efforts have been made to recover spatiotemporal trajectories conditioned on ECG waveforms, the intrinsic principle of EMC that results in various temporal delays remains underexplored in current generative models. To address this limitation, we propose Time2Track, a delay-aware generative model that generates individualized 4D heart motion from same-day ECG recordings by encoding electrophysiology as delay-aware token embeddings. Specifically, a learnable wavelet decomposition is employed to characterize the electromechanical delay, which is then formulated as a physics-based conditioning signal for motion diffusion. Time2Track generates 4D motion with individual-specific kinematics, avoiding the collapse to a population average, i.e., a common failure scenario when diffusing motions whose inter-subject differences are subtle. On $n = 40,577$ subjects from the UK Biobank, our Time2Track accurately synthesizes 4D cardiac trajectories from ECG and preserves inter-subject diversity, a regime in which generic motion diffusion baselines collapse to a population average.

1 Introduction

Electrophysiology signals, including electrocardiogram (ECG) and surface electromyography (sEMG) are the measurement of the coordinated electrical activity of excitable tissues. Electrophysiology data precedes the mechanical actions of human tissues, ruled by a phenomenon known as electromechanical coupling (EMC). Therefore, the heart's depolarization produces the QRS complex, three combined waves (Q, R, and S) representing the main spike in ECG, milliseconds before ventricular contraction, governed by calcium-mediated excitation and contraction coupling [6, 31]. In parallel, motor-unit action potentials in the forearm precede finger joint rotation by a measurable electromechanical delay [7, 23]. Although the challenge of reliable correspondence between electrical dynamics and geometric motion presents a compelling opportunity to faithfully map electrophysiology time series (ETS) to 4D (3D + time) meshes of the heart, the success of ETS-to-4D motion task would allow us to non-invasively recover organ- and limb-level kinesthetics from inexpensive, wearable sensors,

*Corresponding author.

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

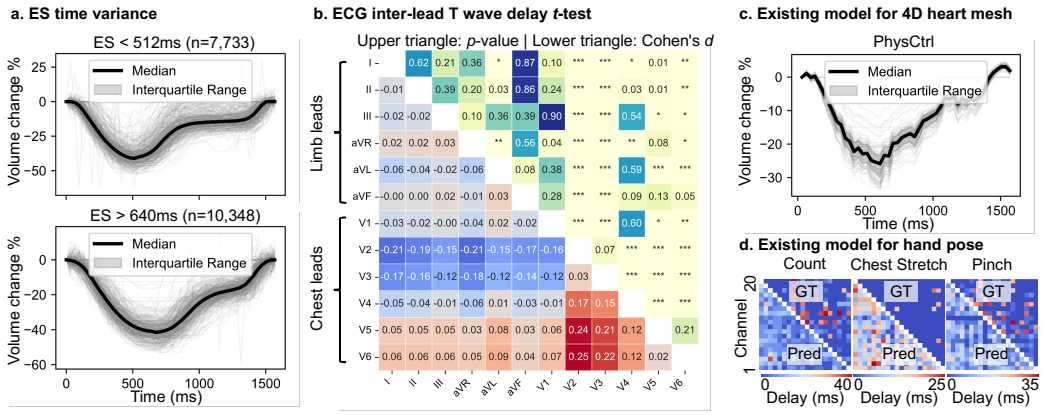

Figure 1: **(a)** The cardiac end-systolic (ES) stage occurred at different time across subjects, where the largest absolute volume change is used to determine the ES stage. **(b)** The ECG inter-lead T wave delay of subjects with ES occurred earlier than 512ms is compared with subjects entering the ES stage later than 640ms using t-test (*: 1e-3$\leq p <$1e-2, **: 1e-4$\leq p <$1e-3, ***: $p <$1e-4), where the delay is mainly observed between chest ECG leads, i.e., V2 and V3. **(c)** PhysCtrl failed on replicating the ES time variance inter-subject. **(d)** Hand pose delay is diverse by gesture, where the delay of ground truth (GT) hand joint angle cannot be replicated by vemg2pose [34] predictions (Pred).

enabling continuous cardiac function monitoring, hand pose tracking, and large-scale phenotyping of movement and pathology.

Despite decades of progress, applications of electrophysiology remain limited for clinical practice. Most pipelines framed the problem into event detection [13], segment classification [3, 32, 2], or scalar regression [34], collapsing a rich spatiotemporal target into low-dimensional labels. The underlying reason is that electrophysiology data are fundamentally misaligned with the motion they encode since the electrical signal leads the mechanical response by a variable delay [6, 7]. Furthermore, the low spatial resolution of multichannel electrodes blurs the geometry of the source [5, 12]. These temporal and spatial misalignments hinder precise end-to-end 4D trajectory prediction.

An effective solution is to leverage recent spatiotemporal generative models for 3D trajectories, such as MeshHeart [29], PhysCtrl [41], and broader human motion diffusion frameworks [38, 52, 35]. However, these models are predominantly conditioned on static attributes and demographics [29, 46], anatomical priors [41], textual descriptions [38, 52], or univariate signals [39], rather than the dynamic, multivariate electrophysiology time series that actually derive motion. One natural attempt is to encode the raw electrophysiology sequence with a temporal backbone and frame it as a condition. However, we argue that this approach is not sufficient due to the ignorance of the multivariate structure of electrophysiology, in which each channel pair exhibits its own temporal delay and causal relationship to the resulting motion, and these relationships vary substantially across subjects and behaviors.

The most supporting physiological evidence is that the timing delay across ECG leads reflects the sequence of regional ventricular activation, and the resulting electrical and mechanical dys-synchrony is closely coupled together in deformation imaging [15]. Under this premise, subject-to-subject variability in inter-channel delays should manifest as corresponding variability in motion timing, i.e., an EMC fundamental that generative models should learn. Figure 1 quantifies this gap on two modalities. For cardiac EMC, we stratify UK Biobank [28] subjects, with 4D meshes derived via established cardiac segmentation and shape-modeling pipelines [4, 49], by the time at which the left ventricle reaches end-systole (ES), which is defined as the moment of largest absolute volume change (Fig. 1a). Such inter-subject variability in end-systolic timing is consistent with long-documented dispersion of systolic time intervals across healthy and diseased populations [47, 18]. Subjects reaching ES before 512ms and those reaching ES after 640ms exhibit significantly different inter-lead T-wave delays, with the effect mainly shown in the chest leads V2 and V3 (Fig. 1b), mirroring the known dispersion of ventricular repolarization captured by regional T-wave morphology [25, 26]. A state-of-the-art (SOTA) physics-conditioned generator, PhysCtrl [41], collapses this inter-subject ES timing variance, producing nearly identical volume-change trajectories regardless of the underlying ECG (Fig. 1c). On the other hand, a parallel pattern emerges for hand kinesthetics. The delay

between sEMG and joint-angle trajectories diverges across gestures, and predictions from vemg2pose [34] fail to reproduce the ground-truth delay distribution (Fig. 1d). Taken together, these observations support our argument that accurate 4D generation from electrophysiology requires a conditioning mechanism that is explicitly delay-aware across various spatial-temporal scales from multivariate ETS, rather than treating it as a generic temporal or multivariate embedding.

To that end, we introduce **Time2Track**, a delay-aware generative model of electromechanical coupling that generates 4D trajectories of individualized heart kinematics from electrophysiological dynamics. Time2Track differentiates from existing works in (1) a delay conditioning mechanism for EMC, developed on the heart and transferred to two sEMG decoders of hand pose in Sec. 4, and (2) a learnable delay modeling of multi-scale wavelets per channel pair. Unlike prior approaches that process electrophysiology through generic temporal backbones or global feature aggregation, Time2Track explicitly disentangles the representation of multivariate signals into distinct morphology and inter-channel delay branches. Existing sequence encoders tend to implicitly average out the fine-grained temporal phase shifts between channels, which explains why models like PhysCtrl and vemg2pose collapse the natural variance of electromechanical coupling into generic, averaged trajectories (as demonstrated in Fig. 1). Time2Track directly solves this spatial and temporal misalignment by explicitly extracting precise temporal offsets across multi-scale wavelets using Fast Fourier Transform Cross-Correlation (FFT-CC). By leveraging these raw, inter-channel time shifts into a high-dimensional, continuous space via a novel, window-aware Fourier feature lag embeddings, we preserve the critical timing differentials that dictate biological motion. Consequently, rather than relying on a black-box correlation, our motion diffusion backbone is conditioned on explicitly quantified physical delays alongside wavelet morphology. This allows the cross-attention mechanism to selectively attend to the exact activation wavefronts that drive mechanical responses, effectively recovering the subject-specific cardiac timing variance.

We summarize our contributions as follows: **(i) Learnable delay tokenization:** We propose a novel mechanism to explicitly extract and tokenize inter-channel delay across multi-scale wavelets from learnable decomposition. By using fast Fourier transform cross-correlation and proposing Fourier feature lag embeddings, these continuous temporal delays are embedded into a high-dimensional latent space. Time2Track preserves the fine-grained phase relationships that dictate individualized motion. **(ii) Delay-aware EMC generator:** We introduce Time2Track, the first delay-aware ETS framework that generates electrophysiology dynamics conditioning 4D motion. By explicitly embedding temporal delay and waveform morphology, it fundamentally aligns electrical signals with mechanical responses. **(iii) SOTA ETS-to-motion generation:** Evaluated on the largest benchmark of ECG-to-motion, over 40,000 combined instances across cardiac [28], Time2Track effectively bridges the gap between 1D biosignals and 4D kinematics, successfully recovering accurate, individual-specific motions from a rest-state prior and multivariate time series.

2 Related Work and Limitations

Deep Learning for Electromechanical Coupling. Modeling the EMC spans both computational physiology and data-driven machine learning. Traditionally, multiscale cardiac models couple ionic channel dynamics with active-stress generation and passive tissue mechanics to simulate the full cardiac cycle from a given action potential [31, 6]. While such physics-based approaches achieve high-fidelity virtual hearts, they depend on subject-specific parameterization via invasive or multimodal imaging and are computationally prohibitive at the population scale. In contrast, data-driven approaches have avoided these requirements at the cost of mechanistic interpretability. For instance, deep neural networks trained on ECG predict scalar indices such as ejection fraction, age, and arrhythmia labels [13, 32, 2], but reduce the 4D trajectory to classification or regression. Crucially, none of the aforementioned methods model the inter-channel delay structure of electrophysiology, nor do they serve as an explicit conditioning variable for predicting the continuous 4D motion of the target anatomy.

Representation Learning for Electrophysiological Signals. CoTAR [50] observes that electrophysiological signals typically originate from centralized biological sources, and replaces the decentralized "all-to-all" self-attention of standard Transformers with a centralized aggregation mechanism built around a learnable core token, as attention along the neural pathways between brain regions does for structural-functional coupling [44, 43]. However, this design does not explicitly capture the temporal relationships between events occurring across different channels. On the other hand, PhysioWave

[8] learns a softmax-weighted combination of wavelet templates across multiple decomposition scales within a BERT-style pretraining framework, yielding morphological tokens that are robust to noise and generalize across ECG and sEMG modalities, much as pretrained foundation models of brain fMRI generalize across cognitive tasks and cohorts [42, 45]. Yet, like CoTAR, PhysioWave encodes multichannel structure through joint attention over all channels, leaving the inter-channel phase relationships entangled, precisely the relationships that carry directional information about the underlying activation wavefront and drive inter-subject variability in kinematics. A broader line of work, Differentiable Digital Signal Processing (DDSP) [10], integrates classical signal-processing operators into end-to-end trainable networks, including parameterized band-pass filters [30] and learnable wavelet networks [20, 11, 33, 22]. These methods, however, target generic feature extraction rather than the modeling of electromechanical coupling. To the best of our knowledge, no prior work extracts inter-channel delays as a structured intermediate representation tied to event localization, nor integrates such delays with adaptive wavelet decomposition.

Conditional Diffusion Models for ECG and 4D Motion. Existing ECG diffusion work targets signal synthesis rather than motion, e.g., SSSD-ECG [1], DiffuSETS [16], ECGTwin [17], and SE-Diff [36]. Human motion diffusion models such as MDM [38], MotionDiffuse [52], PhysCtrl [41], and music-driven approaches like EDGE [39] synthesize articulated body motion conditioned on text descriptions, physical laws or univariate audio signals, but are not designed to accept multivariate electrophysiology as a conditioning input. Cardiac-specific generators, including MeshHeart [29], produce 4D heart mesh trajectories in exclusive conditions on static demographic features and anatomical shape priors rather than the dynamic ECG signal that mechanistically drives the motion. The critical and unaddressed gap is therefore a generative model that (i) accepts multivariate electrophysiology time series as its primary conditioning signal and (ii) outputs a continuous, individualized 4D motion, precisely the contribution of Time2Track.

3 Method

Time2Track takes a resting ECG and a single rest-state heart mesh and produces the individualized 4D trajectory. To prevent the collapse to the population average, we use a learnable multi-scale wavelet decomposition to form morphology and delay tokens for spatio-temporal diffusion conditioning.

Problem Definitions: This work aims to address the problem of trajectory generation in conditions of electrophysiological signals. Given a multivariate electrophysiological signal $\mathbf{e} \in \mathbb{R}^{C\times T}$ comprising C channels, and a subject-specific rest-state representation $\mathbf{p}_0 \in \mathbb{R}^{N\times S}$ (where N denotes the number of control points, with each point containing S features, e.g., for a 3D mesh, $S = 3$), we propose the Time2Track model. This model is designed to recover the trajectory of each control point as it evolves over time, specifically in the following form:

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_F] \in \mathbb{R}^{F\times N\times S}, \mathbf{x}_f \in \mathbb{R}^{N\times S}. \tag{1}$$

Time2Track leverages a Denoising Diffusion Probabilistic Model (DDPM) [14] to model the conditional distribution $p(\mathbf{X}|\mathbf{e}, \mathbf{p}_0)$. Per DDPM, the model consists of forward process, which involves progressively adding Gaussian noise to $\mathbf{X}$ across diffusion time steps $k \in \{1, \ldots, K\}$ to generate $\mathbf{X}^{(k)} = \sqrt{\bar{\alpha}_k}\mathbf{X} + \sqrt{1-\bar{\alpha}_k}\boldsymbol{\epsilon}$, and a denoising network f_θ to predict the denoised trajectory.

$$\hat{\mathbf{X}} = f_\theta\big(\mathbf{X}^{(k)}, k, \mathbf{p}_0, \mathbf{c}(\mathbf{e})\big), \tag{2}$$

where $\mathbf{c}(\mathbf{e}) \in \mathbb{R}^{M\times D}$ denotes conditional tokens generated by our proposed electrophysiology encoder, comprising M tokens, each with a dimensionality of D. The model's training objective combines a motion reconstruction loss with the standard $\boldsymbol{\epsilon}$-prediction loss:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{k,\boldsymbol{\epsilon},(\mathbf{X},\mathbf{e},\mathbf{p}_0)}\left[\left\|\mathbf{X} - f_\theta\big(\mathbf{X}^{(k)}, k, \mathbf{p}_0, \mathbf{c}(\mathbf{e})\big)\right\|_2^2\right]. \tag{3}$$

The core challenge lies in encoding two complementary types of features that are often overlooked in existing downstream tasks: (i) **Morphology**, specifically the shape characteristics of multiple transient waveforms across different scales, and (ii) **inter-channel delays**, referring to the temporal offsets between signals from different channels, which carry directional information regarding the underlying activation wavefront and the inter-subject difference of kinematic characteristics as illustrated in Fig. 1. This section focuses on our methods of electrophysiology time series encoding.

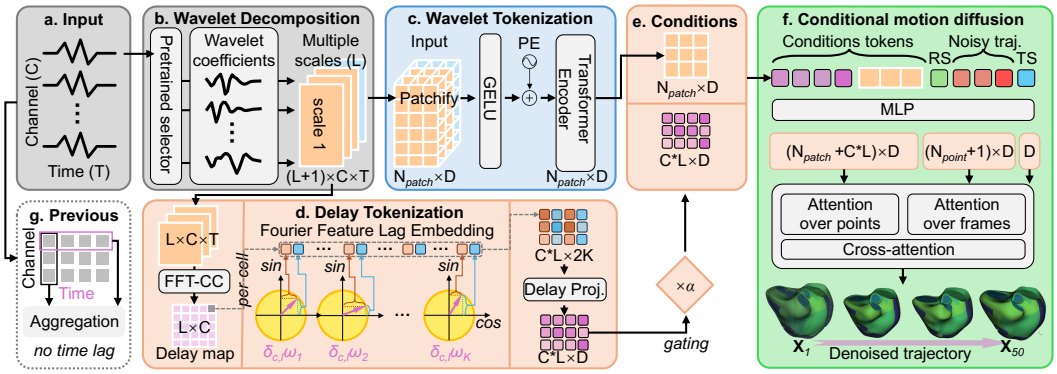

Figure 2: Time2Track framework. **(a)** The multivariate time series input. **(b)** The learnable wavelet decomposition produces the wavelet coefficients, $\mathbf{W} \in \mathbb{R}^{(L+1)\times C\times T_{\text{pad}}}$. The orange matrices denote detail scales blue is an approximation scale. **(c)** The wavelet branch first patchifies $\mathbf{W}$ to a sequence of windowed time series, and then it encodes tokens via Transformer encoder along with a positional embedding (PE). **(d)** The delay branch takes the detail scales $\mathbf{W}_{:L}$ and extracts temporal shift per (channel, scale) as a delay map $\Delta \in \mathbb{R}^{L\times C}$, followed by a conversion from delay value to a $2K$-long vector for tokenization. **(e)** Conditions of the motion diffusion are concatenated by wavelet and delay tokens, where delay tokens are gated by α. **(f)** Conditional motion diffusion generates a denoised heart mesh trajectory using a spatial-temporal Transformer given wavelet and delay tokens. **(g)** Although previous approaches encode aggregated features across channels and time, Time2Track produces delay tokens for the motion with only subtle and individual-specific differences.

3.1 Delay Aware Electrophysiology Encoder

The workflow of Time2Track is outlined in Fig. 2. It consists of two branches for the two aforementioned challenges, starting from (a) the multi-channel input and (b) a learnable multi-scale wavelet decomposition module. The morphology branch is mainly derived from PhysioWave [8], given the pretrained wavelet decomposition for ECG and EMG to produce robust morphology tokens in (c). The delay branch runs over decomposed wavelets and learns on the delay tokenization in (d). In the final, two branches are merged in (e) and applied in (f) for the conditional motion diffusion.

(a) Input. The raw signal is a z-normalized, C-channel signal segment $\mathbf{e} \in \mathbb{R}^{C\times T}$. For the morphology branch, $\mathbf{e}$ is zero-padded to a length of $T_{\text{pad}} = \lceil T/P \rceil P$, where P denotes the length of a temporal block. Whereas for the delay branch, we utilize the unpadded raw $\mathbf{e}$ to avoid spurious correlations introduced by zero-padding.

(b) Wavelet decomposition. $\mathbf{e}$ is then mapped by a pretrained wavelet decomposition model, PhysioWave, to a multi-scale spectrogram. The selector of PhysioWave mixes a bank of wavelet template $\{\psi_m\}_{m=1}^{M_\psi}$ via a softmax-weighted convex combination per scale, resulting in a L-level decomposition of detail coefficients $\{\mathbf{d}^{(\ell)}\}_{\ell=1}^{L}$ and an approximation $\mathbf{a}^{(L)}$. We stack them along the channel axis and denote it as $\mathbf{W}$ for the following wavelet tokenization:

$$\mathbf{W} = \text{Concat}\big(\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(L)}, \mathbf{a}^{(L)}\big) \in \mathbb{R}^{(L+1)\times C\times T_{\text{pad}}}. \tag{4}$$

(c) Morphology branch (wavelet tokenization). $\mathbf{W}$ is first patchified into $N_{\text{patch}} = (L+1)C\lceil T_{\text{pad}}/P \rceil$ time blocks (1D patches) over (channel, scale) using a convolutional layer Patch, followed by a GELU non-linearity and 2D positional embedding $\mathbf{PE}_{\text{2d}}$. Then, a stack of pre-norm Transformer blocks $\mathcal{T}_{\text{morph}}$ produces N_{patch} tokens

$$\mathbf{H}_{\text{morph}} = \mathcal{T}_{\text{morph}}\big(\text{GELU}(\text{Patch}(\mathbf{W})) + \mathbf{PE}_{\text{2d}}\big) \in \mathbb{R}^{N_{\text{patch}}\times D'}, \tag{5}$$

which we project to the latent width via a linear map: $\mathbf{T}_{\text{morph}} = MLP_{\text{morph}}(\mathbf{H}_{\text{morph}}) \in \mathbb{R}^{N_{\text{patch}}\times D}$.

(d) Delay branch (delay tokenization). The delay branch extracts the temporal shift between every channel and a reference channel r, at every detail scale of wavelet from $\mathbf{W}$, where only detail bands are used and denoted by $W_c^{(\ell)} \in \mathbb{R}^T$ with $\ell \in \{1, \ldots, L\}$ and channel $c \in \{1, \ldots, C\}$. The fast

Fourier transform cross correlation (FFT-CC) is utilized as shown in Fig. 3 to measure the delay. It first gets cross-correlation

$$R_c^{(\ell)}[\tau] = \frac{\mathcal{F}^{-1}\big(\widetilde{W}_c^{(\ell)} \odot \overline{\widetilde{W}_r^{(\ell)}}\big)[\tau]}{\|W_c^{(\ell)}\|_2 \|W_r^{(\ell)}\|_2 + \varepsilon}, \widetilde{W} = \mathcal{F}(W), \tag{6}$$

where $\mathcal{F}$ denotes FFT, and $\overline{\widetilde{W}_r^{(\ell)}}$ denotes element-wise complex conjugate of the reference spectrum, so that the element-wise multiplication of $\widetilde{W}_c^{(\ell)}$ and $\overline{\widetilde{W}_r^{(\ell)}}$ is the cross-spectrum and the inverse fast Fourier transform (IFFT) is the circular cross-correlation by the Wiener-Khinchin theorem [24].

We restrict the soft-argmax to a physically plausible window $\tau \in [-\tau_{\max}, \tau_{\max}]$ depending on the data modality, resulting in a temperature-controlled soft-argmax as follows:

$$\hat{\tau}_c^{(\ell)} = \sum_{\tau=-\tau_{\max}}^{\tau_{\max}} \tau \frac{\exp\big(\beta_\ell R_c^{(\ell)}[\tau]\big)}{\sum_{\tau'} \exp\big(\beta_\ell R_c^{(\ell)}[\tau']\big)}, \beta_\ell = \mathrm{softplus}(\beta_\ell^{\mathrm{raw}}) + \epsilon. \tag{7}$$

where $\beta_\ell^{\mathrm{raw}} \in \mathbb{R}$ is a learnable scalar parameter. Therefore, inter-channel delay can be defined by $\delta_c^{(\ell)} = \hat{\tau}_c^{(\ell)} - \hat{\tau}_r^{(\ell)}$, which obviously exhibits shift invariance in $\mathbf{e}$. Afterward, the resulting delay map $\mathbf{\Delta} \in \mathbb{R}^{C \times L}$ is tokenized via a Fourier feature lag embedding approach.

Fourier feature lag embedding. Continuous lags are notoriously hard for MLPs to extract features. We therefore lift each $\delta_c^{(\ell)}$ to a sinusoidal embedding analogous to neural radiance field (NeRF) positional encodings [21], but window-aware: We use K log-spaced angular frequencies tied to the search window, $\omega_k = \frac{2^k \pi}{\tau_{\max}}$, $k = 0, \ldots, K-1$, yielding a $2K$-dimensional feature

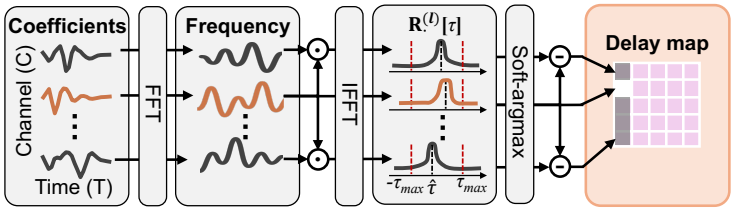

Figure 3: The workflow of fast Fourier transform cross correlation (FFT-CC) to extract the delay map from decomposed wavelet coefficients, where the orange channel denotes the reference channel, $\odot$ denotes element-wise multiplication, and $R_c^{(\ell)}$ refers to Eq. 6. The reference channel is applied $\odot$ and $\ominus$ to itself.

$$\boldsymbol{\phi}\big(\delta_c^{(\ell)}\big) = \big[\sin(\omega_0\delta), \cos(\omega_0\delta), \ldots, \sin(\omega_{K-1}\delta), \cos(\omega_{K-1}\delta)\big]. \tag{8}$$

The sine/cosine clock plot in Fig. 2d vividly illustrates this multi-resolution lifting process. As the dimensionality of inter-channel time lag is lifted, the delay projector can robustly map each $\boldsymbol{\phi}(\delta_c^{(\ell)}) \in \mathbb{R}^{2K}$ to $\mathbb{R}^D$. Note that channel and scale embeddings, which are omitted in Fig. 2d for clarity, $\mathbf{u}_c, \mathbf{v}_\ell \in \mathbb{R}^D$ are also accumulated to every token of the delay map during the delay projection

$$\mathbf{T}_{\mathrm{delay}} = \alpha \cdot \mathcal{T}_{\mathrm{delay}}\big([g_\phi\big(\boldsymbol{\phi}(\delta_c^{(\ell)})\big) + \mathbf{u}_c + \mathbf{v}_\ell]_{c,\ell}\big) \in \mathbb{R}^{C \cdot L \times D}, \tag{9}$$

where the learnable scalar parameter $\alpha \in \mathbb{R}$ is zero initialized at the start of training, so that the encoder is stable at the beginning to just fine-tune the wavelet decomposition.

(e) Condition tokens. The morphology and delay tokens are concatenated as the condition tokens, $\mathbf{c}(\mathbf{e}) = \mathrm{Concat}\big(\mathbf{T}_{\mathrm{delay}}, \mathbf{T}_{\mathrm{morph}}\big) \in \mathbb{R}^{(C \cdot L + N_{\mathrm{patch}}) \times D}$, in which the type and ordering are retrievable through the channel, scale, and positional embeddings. We deliberately avoid pooling either branch so that cross-attention in Fig. 2f can attend selectively to morphological wavelets or to a specific delay token at channel c, scale ℓ.

3.2 Electrophysiology Conditional Motion Diffusion

The electrophysiology conditional Motion Diffusion Model (MDM) shown in Fig. 2f is a Spatio-Temporal Transformer, the same as general MDMs [38, 41], for kinematic trajectory. First, the rest-state (RS) representation $\mathbf{p}_0$ (green token in Fig. 2f) is prepended to $\mathbf{X}$ as $\widetilde{\mathbf{X}} \in \mathbb{R}^{(F+1) \times N \times S}$ so that the initial pose is attended. Then, it is mapped to the latents of 4D motion denoted by

$\mathbf{Z} \in \mathbb{R}^{(F+1)\times N\times D}$ via an MLP projector. The backbone interleavedly proceeds spatial self-attention over the N points within a frame and temporal self-attention over the $F+1$ frames for the same point, with the diffusion timestep (TS, blue token in Fig. 2f) injected via Adaptive Layer Normalization (AdaLN), updating $\mathbf{Z}$ every layer. Conditioning on the input signal is implemented by cross-attention

$$\mathbf{Z}'_{f,n} = \mathbf{Z}_{f,n} + \mathrm{CrossAttn}\big(\mathbf{Z}_{f,n}, \mathbf{c}(\mathbf{e})\big). \tag{10}$$

After L_{dit} layers, an MLP head (omitted in Fig. 2) reads off the predicted trajectory offset $\hat{\mathbf{X}}$ with respect with the RS slot

$$\hat{\mathbf{X}} = \hat{\mathbf{X}}_{1:F} + \mathbf{p}_0, \tag{11}$$

which improves convergence by enforcing the trivial trajectory as a cheap prior. At inference, we run DDIM sampling [37] for K_{infer} steps and apply classifier-free guidance.

4 Experimental Results

Datasets. The experiments of this work involve two applications, ECG conditional heart motion generation using UKB [28]. UKB contains large subject and kinematic variability that has demonstrated EMC diversity in Fig. 1, resulting in a challenging task for general motion diffusion models and existing electrophysiology encoders without inter-channel delay consideration. UKB has 53,224 subjects that span the age range from adults to the elderly. The 4D heart mesh is extracted from ECG-gated cardiac MRI for one cycle, and quality controlled, released by [40]. and it is represented by the 388×3 control points. After pairing with resting-state 12-lead ECG (500Hz) captured at the same visit as MRI and filtering samples with missing frames, there are $n = 40,577$ ECG-mesh pairs. Electrophysiology-motion pairs are time aligned, and the train:validation ratio is 80:20%. Since the protocol has no held-out test split, every checkpoint is selected by the training objective itself, the one-step DDPM denoising loss, so that no fidelity or diversity metric enters model selection, and all encoders share one training configuration without hyperparameter search.

Baselines and implementations. To our knowledge, no prior diffusion model conditions 4D motion on multi-variate electrophysiology signal as reviewed in Section 2. Therefore, we hold the diffusion backbone based on a general-purpose state-of-the-art (SOTA) MDM, PhysCtrl, and vary the electrophysiology encoder, so that experimental results isolate the contribution of the conditioning representation. The diffusion backbone is a Spatial-Temporal Transformer with 8 layers, 256 latent channels, 8 conditioning tokens, 50 DDIM inference steps, etc., as the default settings of PhysCtrl. We also evaluated MDM to provide a sanity check for a generic motion-diffusion approach on this task. Encoders chosen for ECG-mesh data: **(i) CNN**: a vanilla 1D-CNN encoder over the 12-lead waveform that emits 8 conditioning tokens, serving as a minimal baseline. **(ii) TeCh** [50]: a temporal-channel transformer that used 2-layer and 2-layer centralized self-attention over channel and time, respectively, producing an SOTA fused ECG token. **(iii) PhysioWave** [8]: a BERT-style wavelet foundation model, mixed 4 types of wavelets, embed dim 384, depth 8, 12 heads, and initialized from publicly released ECG pretrained weights. **(iv) ShiftInvariant** [9]: a shift invariant model capturing inter-lead timing implicitly without any external annotation. **(v) Delay-T**: the CNN encoder augmented with an explicit 12×12 inter-lead T-peak delay matrix that is encoded by a per-row MLP (hidden 128) into 12 additional conditioning tokens, providing direct supervision of a manually selected delay type. **(vi) Delay-QRS**: identical to Delay-T, but the delay of the QRS complex is manually selected. Table 6 in Appendix A.4 summarizes the architecture, the handling of time lag and the size of every encoder.

Metrics. The measurement of experimental results focuses on both global and local structures, trends, and absolute value of location, speed, and surface area, as well as the similarity of inter-subject diversity between the predicted and true 4D motion. We contrast predicted (Pred) motion against ground truth (GT) along complementary axes as follows: **(i) Global trend** is measured by the per-subject **RV coefficient** and Rooted Mean Squared Error (**RMSE**) between the raw Pred and GT trajectories, capturing whether the model recovers the dominant systolic-diastolic pattern across the cycle, which is ruled by EMC. **(ii) Local trend.** After cubic detrending of each curve to suppress the global trend, **PAR=**$(max(\hat{\text{Pred}}) - min(\hat{\text{Pred}}))/(max(\hat{\text{GT}}) - min(\hat{\text{GT}}))$ reports the ratio between peak-to-peak magnitudes of detrended predicted ($\hat{\text{Pred}}$) and GT ($\hat{\text{GT}}$) trajectories. PAR quantifies 'does it put local peaks at the right size'. **(iii) Inter-subject diversity** is the main measurement in this work, designed to quantify whether a model collapses to a population mean when the inter-subject

Table 1: Whole-heart 3D vertex trajectory metrics (Pred vs. GT). Inter-subject diversity columns are reported as $\widehat{\text{Pred}}/\widehat{\text{GT}}$ ratios, with $\pm$ the SD over the 10,000 paired subject bootstrap resamples described in Appendix A.5.

Method	RV Coefficient↑	RMSE↓	PAR↑	$\lvert\text{APCR} - 1\rvert \downarrow$	$\lvert\text{APDR} - 1\rvert \downarrow$
CNN	0.634±0.150	0.130±0.046	0.466±0.106	0.241±0.058	0.151±0.007
TeCh	0.636±0.153	**0.129±0.045**	0.401±0.098	0.356±0.065	0.170±0.012
PhysioWave	0.634±0.174	0.172±0.074	0.868±0.090	0.044±0.037	0.079±0.006
ShiftInvariant	0.558±0.168	0.144±0.052	0.729±0.249	0.452±0.078	0.305±0.009
Delay-T	0.619±0.177	0.158±0.057	0.772±0.082	0.042±0.046	0.091±0.008
Delay-QRS	0.589±0.179	0.166±0.061	0.894±0.145	0.322±0.049	0.025±0.006
Ours	**0.637±0.176**	0.151±0.069	**0.896±0.130**	**0.034±0.042**	**0.019±0.008**

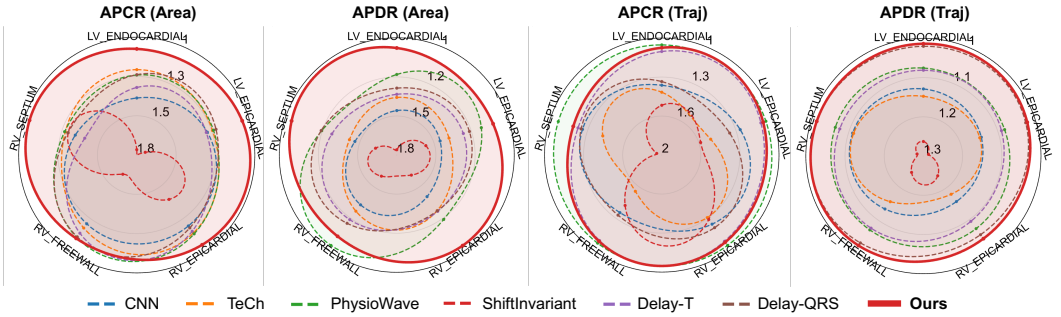

Figure 4: Radar charts of inter-subject diversity metrics, including Average Pairwise Coefficient Ratio (APCR) and Average Pairwise Distance Ratio (APDR) evaluated on mesh face area (Area) and vertex trajectory (Traj). The metrics are computed across five distinct anatomical regions of the heart preprocessed by [40], where RV/LV denotes Right/Left Ventricular. The outermost line of each radar represents the ideal ratio of 1.0, meaning the predicted diversity perfectly matches the ground truth.

differences in 4D motion are subtle. We report the average pairwise distance ratio (**APDR**) and average pairwise coefficient ratio (**APCR**) on the 3D vertex trajectory (**Traj**), the 1D mesh face area (**Area**), and the 1D speed of the vertex trajectory (**Speed**). For each metric, we compute the subject-to-subject distance or similarity coefficient (RV for 3D Traj, Pearson for 1D) on detrended predicted and GT curves, with detrending applied to remove the dominant global heartbeat pattern. APDR and APCR are then the ratios of the predicted to the GT values. Ratios far from one indicate collapse to a subject-invariant amplitude or slope. These diversity metrics are essential because UKB exhibits strong inter-subject variability (Fig. 1a), and a model that produces nearly identical motions for every subject can otherwise score deceptively well on global-trend metrics.

Quantitative Results: Table 1 presents the quantitative evaluation of the predicted whole-heart 3D vertex trajectories against the GT. While centralized encoders like CNN and TeCh achieve competitive RMSE scores (0.130 and 0.129, respectively), their significantly low PAR values (0.466 and 0.401) indicate an inability to match peak-to-peak local magnitudes. Furthermore, they exhibit high errors in inter-subject diversity metrics ($\lvert\text{APCR} - 1\rvert$ and $\lvert\text{APDR} - 1\rvert$), reflecting a severe collapse where the model predicts an averaged heart motion for all subjects without the subtle but individual-specific differences. In contrast, Time2Track achieves the best performance in capturing both the global structure and local trends, yielding the highest RV Coefficient (0.637) and PAR (0.896). Crucially, our method dominates in preserving inter-subject diversity, obtaining the lowest $\lvert\text{APCR} - 1\rvert$ (0.034) and $\lvert\text{APDR} - 1\rvert$ (0.019), which are nearly $10\times$ better than vanilla baselines. While explicitly supervising certain delays (Delay-T, Delay-QRS) or using foundation models (PhysioWave) improves diversity over the CNN baseline, they still fall short of our approach with the learnable delay.

These whole-heart observations are further supported by the fine-grained regional diversity analysis in the radar charts (Fig. 4). Across five distinct anatomical regions and across both face area (Area) and vertex trajectory (Traj) metrics, our method (solid red line) consistently stays closest to the ideal ratio of 1.0. Almost all baseline methods systematically underestimate this diversity, drawing inward toward the center of the radar charts, confirming that they tend to generate smoothed and

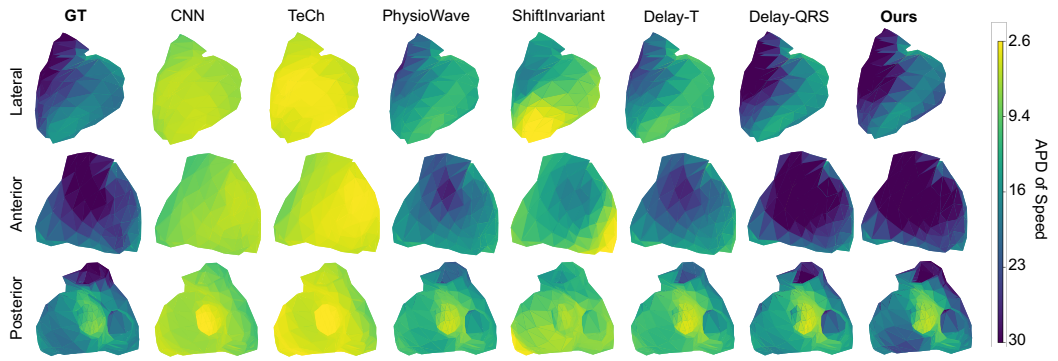

Figure 5: Qualitative visualization of inter-subject diversity, measured by the Average Pairwise Distance (APD) of vertex speed mapped onto the 3D heart meshes. The meshes are shown from three anatomical perspectives: Lateral, Anterior, and Posterior. The colormap indicates the mean of per-vertex APD across the validated subjects, where the prediction is expected to reproduce GT.

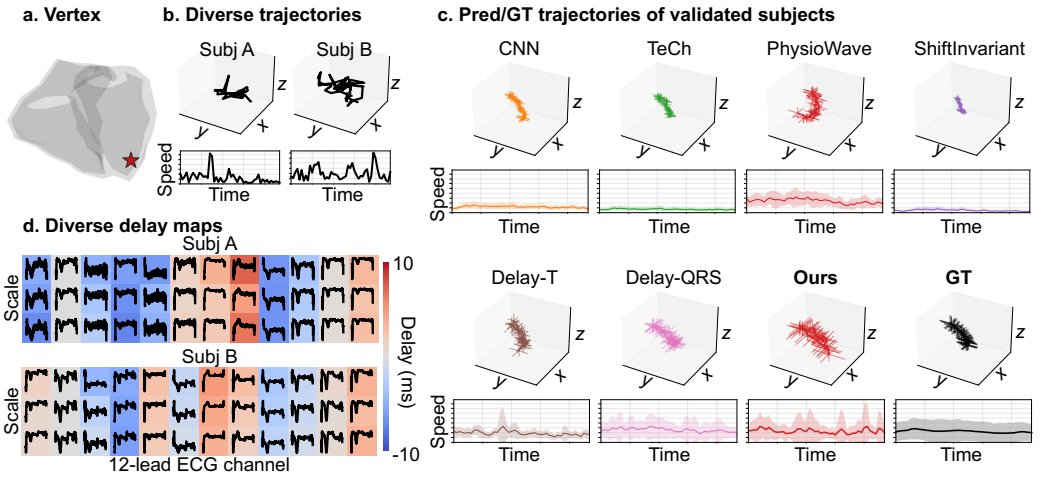

Figure 6: Qualitative comparison of trajectories and delays. **(a)** Randomly selected vertex on the right ventricle used here. **(b)** GT 3D trajectories and speed profiles for two subjects show high natural kinematic diversity. **(c)** Predicted 3D trajectories and speed distributions (mean curve and variance bar) across the validation set. Baselines suffer from severe mode collapse, predicting near-mean motions with near-zero variance. Conversely, Ours accurately recovers the spatial spread and speed variance, closely matching GT. **(d)** Extracted multi-scale inter-channel delay maps with wavelet coefficient curves. These subject-specific temporal shifts drive our model's ability to generate personalized motions instead of collapsing to the average.

averaged cardiac motions. The ablations in Appendix A.3 show that removing the delay branch raises $|\text{APDR} - 1|$ from 0.019 to 0.079 and that FFT-CC and the lag embedding outperform their simpler replacements.

Qualitative Results: Fig. 5 visualizes inter-subject diversity by mapping the Average Pairwise Distance (APD) of vertex speed onto the 3D heart meshes. In contrast, baseline encoders (CNN, TeCh, ShiftInvariant) produce low-variance, yellowish surfaces, failing entirely to capture patient-specific speed dynamics. While PhysioWave and explicit delay encoders show slight improvements, Time2Track is the best at reproducing the spatially heterogeneous APD distribution of the GT.

Furthermore, Fig. 6 highlights trajectory distributions for a randomly selected vertex on the right ventricle. As shown in Fig. 6b, true subject motions (e.g., Subj A vs. Subj B) exhibit highly diverse 3D paths and speed profiles. In Fig. 6c, baseline models collapse to a dense, near-mean trajectory bundle with nearly flat, zero-variance speed distributions. In contrast, Time2Track recovers wide variance bars and larger shade in speed plots, demonstrating a close mirror of GT. As illustrated in

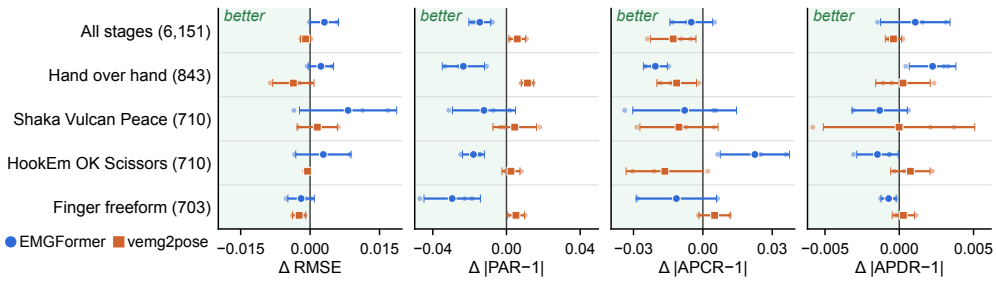

Figure 7: Application to EMG2POSE [34] dataset by adding delay tokens to the sEMG decoders EMGFormer [48] and vemg2pose [34] on the official test split. Dots show three seeds, and bars their mean ± SD.

Fig. 6d, this superior generative capability is driven by our extracted multi-scale inter-channel delay maps.

Pathology and Robustness: We have compared the 981 validation subjects with a cardiac ICD-10 code with as many randomly sampled healthy subjects. Table 2 shows that Time2Track has the smallest RMSE change and the only decrease of $|\text{APCR} - 1|$, which rises by 0.034 for CNN and 0.064 for TeCh. The robustness tests in Appendix A.6 show little effect of electrode noise and τ_{max}, whereas dropping lead II raises $|\text{APCR} - 1|$.

Table 2: Diseased minus matched healthy subjects ($n = 981$, lower is better).

Δ	RMSE	$\lvert\text{PAR} - 1\rvert$	$\lvert\text{APCR} - 1\rvert$	$\lvert\text{APDR} - 1\rvert$
Ours	**+0.001**	+0.009	**−0.025**	+0.004
CNN	+0.005	**−0.007**	+0.034	−0.001
TeCh	+0.005	+0.008	+0.064	**−0.005**

ECG as the Only Input: Replacing the subject's rest mesh $\mathbf{p}_0$ with the population average $\hat{\mathbf{p}}_0$ at inference, without retraining, leaves a resting 12-lead ECG as the only input. Table 3 shows that the cost falls on fidelity (RV −0.046, RMSE +0.031) rather than on diversity.

Table 3: Own ($\mathbf{p}_0$) vs. population ($\hat{\mathbf{p}}_0$) rest mesh.

	$\mathbf{p}_0$	$\hat{\mathbf{p}}_0$	Δ
RV↑	0.637	0.591	−0.046
PAR → 1	0.896	0.867	−0.029
$\lvert\text{APCR} - 1\rvert$	0.034	0.052	+0.018
$\lvert\text{APDR} - 1\rvert$	0.019	0.036	+0.017
RMSE↓	0.151	0.182	+0.031

Transfer of Delay Tokens to sEMG: We inject the delay tokens into the sEMG hand pose decoders vemg2pose [34] and EMGFormer [48], both detailed in Appendix A.8 and trained with three seeds on the first gesture cycle of each emg2pose session (6,151 test sessions, 178 users). Fig. 7 shows that delay tokens lower $|\text{PAR}-1|$ of EMGFormer in every subset and $|\text{APCR}-1|$ of both decoders in four of five subsets, while RMSE and $|\text{APDR}-1|$ change by less than 0.01 and $|\text{PAR} - 1|$ of vemg2pose rises by at most 0.011.

5 Conclusions and Limitations

In this work, we introduced Time2Track, a delay-aware generative model of electromechanical coupling that generates individualized 4D cardiac kinematics directly from multivariate electrophysiology time series. By explicitly modeling learnable inter-channel temporal delays through FFT cross-correlation and Fourier feature lag embeddings, our method effectively overcomes the severe collapse observed in generic motion diffusion models with various existing ECG encoders, successfully recovering natural inter-subject kinematic diversity. In this regard, Time2Track enables two downstream applications, namely 4D motion assessment from a resting ECG with either the rest mesh of an existing cardiac MR scan or the population-average mesh in place of the ECG-gated cine acquisition, and sEMG decoding of hand pose with the same delay tokens. Despite these advances, our approach has three limitations. First, it depends on the searching window size τ_{max} for cross-correlation, which may restrict its applicability in extreme cases with highly abnormal electromechanical delays. Second, the UKB imaging cohort is about 90% healthy with registry-coded diagnoses, so severe or acute pathology remains untested. Third, the 388 control points are coarser than the ventricular wall thickness, so regional wall-motion abnormalities remain beyond our model and metrics. Future work will focus on higher-resolution representations, validation in clinical cohorts, and real-time motion tracking in severe cardiovascular and neuromuscular disorders.

Acknowledgments and Disclosure of Funding

This work was supported by the National Institutes of Health (AG091653, AG068399, AG084375) and the Foundation of Hope.

Competing Interests

The authors declare no competing interests.

References

[1] Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based conditional ecg generation with structured state space models. *Computers in biology and medicine*, 163:107115, 2023.

[2] Zachi I. Attia, Suraj Kapa, Francisco Lopez-Jimenez, Paul M. McKie, Dorothy J. Ladewig, Gaurav Satam, Patricia A. Pellikka, Maurice Enriquez-Sarano, Peter A. Noseworthy, Thomas M. Munger, Samuel J. Asirvatham, Christopher G. Scott, Rickey E. Carter, and Paul A. Friedman. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature Medicine*, 25(1):70–74, 2019.

[3] Manfredo Atzori, Matteo Cognolato, and Henning Müller. Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands. *Frontiers in Neurorobotics*, 10:9, 2016.

[4] Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M. Lee, Nay Aung, Elena Lukaschuk, Mihir M. Sanghvi, Filip Zemrak, Kenneth Fung, Jose Miguel Paiva, Valentina Carapella, Young Jin Kim, Hideaki Suzuki, Bernhard Kainz, Paul M. Matthews, Steffen E. Petersen, Stefan K. Piechnik, Stefan Neubauer, Ben Glocker, and Daniel Rueckert. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 20(1):65, 2018.

[5] Laura R Bear, Ian J LeGrice, Gregory B Sands, Nigel A Lever, Denis S Loiselle, David J Paterson, Leo K Cheng, and Bruce H Smaill. How accurate is inverse electrocardiographic mapping? a systematic in vivo evaluation. *Circulation: Arrhythmia and Electrophysiology*, 11(5):e006108, 2018.

[6] Donald M. Bers. Cardiac excitation-contraction coupling. *Nature*, 415(6868):198–205, 2002.

[7] P. R. Cavanagh and P. V. Komi. Electromechanical delay in human skeletal muscle under concentric and eccentric contractions. *European Journal of Applied Physiology and Occupational Physiology*, 42(3):159–163, 1979.

[8] Yanlong Chen, Mattia Orlandi, Pierangelo Maria Rapa, Simone Benatti, Luca Benini, and Yawei Li. Physiowave: A multi-scale wavelet-transformer for physiological signal representation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[9] Berken Utku Demirel and Christian Holz. Shifting the paradigm: A diffeomorphism between time series data manifolds for achieving shift-invariancy in deep learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[10] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*, 2020.

[11] Gaëtan Frusque and Olga Fink. Learnable wavelet packet transform for data-adapted spectrograms. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3119–3123. IEEE, 2022.

[12] Carina M Germer, Dario Farina, Leonardo A Elias, Stefano Nuccio, François Hug, and Alessandro Del Vecchio. Surface emg cross talk quantified at the motor unit population level for muscles of the hand, thigh, and calf. *Journal of applied physiology*, 131(2):808–820, 2021.

[13] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, 2019.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[15] Pavel Jurak, Josef Halamek, Jaroslav Meluzin, Filip Plesinger, Tereza Postranecka, Jolana Lipoldova, Miroslav Novak, Vlastimil Vondra, Ivo Viscor, Ladislav Soukup, Petr Klimes, Petr Vesely, Josef Sumbera, Karel Zeman, Roshini S. Asirvatham, Jason Tri, Samuel J. Asirvatham, and Pavel Leinveber. Ventricular dyssynchrony assessment using ultra-high frequency ECG technique. *Journal of Interventional Cardiac Electrophysiology*, 49(3):245–254, 2017.

[16] Yongfan Lai, Jiabo Chen, Qinghao Zhao, Deyun Zhang, Yue Wang, Shijia Geng, Hongyan Li, and Shenda Hong. Diffusets: 12-lead ecg generation conditioned on clinical text reports and patient-specific information. *Patterns*, 6(10), 2025.

[17] Yongfan Lai, Bo Liu, Xinyan Guan, Qinghao Zhao, Hongyan Li, and Shenda Hong. Ecgtwin: Personalized ecg generation using controllable diffusion model. *arXiv preprint arXiv:2508.02720*, 2025.

[18] Richard P. Lewis, Stanley E. Rittgers, William F. Forester, and Harisios Boudoulas. A critical review of the systolic time intervals. *Circulation*, 56(2):146–158, 1977.

[19] Steve Meek and Francis Morris. ABC of clinical electrocardiography: Introduction. I: Leads, rate, rhythm, and cardiac axis. *BMJ*, 324(7334):415–418, 2002.

[20] Gabriel Michau, Gaetan Frusque, and Olga Fink. Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series. *Proceedings of the National Academy of Sciences*, 119(8):e2106598119, 2022.

[21] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[22] LIU Minhao, Ailing Zeng, LAI Qiuxia, Ruiyuan Gao, Min Li, Jing Qin, and Qiang Xu. T-wavenet: A tree-structured wavelet neural network for time series signal analysis. In *International conference on learning representations*, 2021.

[23] A. Nordez, T. Gallot, S. Catheline, A. Guével, C. Cornu, and F. Hug. Electromechanical delay revisited using very high frame rate ultrasound. *Journal of Applied Physiology*, 106(6):1970–1975, 2009.

[24] Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

[25] Tobias Opthof, Ruben Coronel, Francien J. G. Wilms-Schopman, Alexei N. Plotnikov, Iryna N. Shlapakova, Peter Danilo, Michael R. Rosen, and Michiel J. Janse. Dispersion of repolarization in canine ventricle and the electrocardiographic T wave: Tp-e interval does not reflect transmural dispersion. *Heart Rhythm*, 4(3):341–348, 2007.

[26] Ragesh Panikkath, Kyndaron Reinier, Audrey Uy-Evanado, Carmen Teodorescu, Jason Hattenhauer, Ronald Mariani, Karen Gunson, Jonathan Jui, and Sumeet S. Chugh. Prolonged Tpeak-to-Tend interval on the resting ECG is associated with increased risk of sudden cardiac death. *Circulation: Arrhythmia and Electrophysiology*, 4(4):441–447, 2011.

[27] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[28] Steffen E. Petersen, Paul M. Matthews, Jane M. Francis, Matthew D. Robson, Filip Zemrak, Redha Boubertakh, Alistair A. Young, Sarah Hudson, Peter Weale, Steve Garratt, Rory Collins, Stefan Piechnik, and Stefan Neubauer. UK Biobank's cardiovascular magnetic resonance protocol. *Journal of Cardiovascular Magnetic Resonance*, 18(1):8, 2016.

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

[30] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)*, pages 1021–1028. IEEE, 2018.

[31] F. Regazzoni, M. Salvador, P. C. Africa, M. Fedele, L. Dede', and A. Quarteroni. A cardiac electromechanical model coupled with a lumped-parameter model for closed-loop blood circulation. *Journal of Computational Physics*, 457:111083, 2022.

[32] Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira Jr., Thomas B. Schön, and Antonio Luiz P. Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11:1760, 2020.

[33] David W Romero, Erik J Bekkers, Jakub M Tomczak, and Mark Hoogendoorn. Wavelet networks: Scale-translation equivariant learning from raw time-series. *arXiv preprint arXiv:2006.05259*, 2020.

[34] Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Shangchen Han, Rohin Bhasin, Yujun Cai, Peter Walkington, Anuoluwapo Bolarinwa, Robert Wang, Nathan Danielson, Josh Merel, Eftychios Pnevmatikakis, and Jesse Marshall. emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation. In *Advances in Neural Information Processing Systems 37 (NeurIPS) Track on Datasets and Benchmarks*, 2024.

[35] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations (ICLR)*, 2024.

[36] Dongseok Shim and H Jin Kim. Sediff: Structure extraction for domain adaptive depth estimation via denoising diffusion models. In *European Conference on Computer Vision*, pages 37–53. Springer, 2024.

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[38] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023.

[39] Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. EDGE: Editable dance generation from music. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 448–458, 2023.

[40] Devran Ugurlu, Shuang Qian, Elliot Fairweather, Charlene Mauger, Bram Ruijsink, Laura Dal Toso, Yu Deng, Marina Strocchi, Reza Razavi, Alistair Young, et al. Cardiac digital twins at scale from mri: open tools and representative models from˜ 55000 uk biobank participants. *Plos one*, 20(7):e0327158, 2025.

[41] Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. Physctrl: Generative physics for controllable and physics-grounded video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[42] Ziquan Wei, Tingting Dan, Tianlong Chen, and Guorong Wu. BrainMoE: Cognition joint embedding via mixture-of-expert towards robust brain foundation model. In *Advances in Neural Information Processing Systems*, volume 38, pages 4541–4563, 2025.

[43] Ziquan Wei, Tingting Dan, Jiaqi Ding, Paul J. Laurienti, and Guorong Wu. NeuroDetour: A neural pathway transformer for uncovering structural-functional coupling mechanisms in human connectome. *Medical Image Analysis*, 109:103931, 2026.

[44] Ziquan Wei, Tingting Dan, Jiaqi Ding, and Guorong Wu. NeuroPath: A neural pathway transformer for joining the dots of human connectomes. In *Advances in Neural Information Processing Systems*, volume 37, pages 67826–67849, 2024.

[45] Ziquan Wei, Tingting Dan, and Guorong Wu. Large connectome model: An fMRI foundation model of brain connectomes empowered by brain-environment interaction in multitask learning landscape. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 2155–2163, 2026.

[46] Ziquan Wei, Tingting Dan, and Guorong Wu. Marrying generative model of healthcare events with digital twin of social determinants of health for disease reasoning. In *International Conference on Machine Learning*, 2026. arXiv:2605.09771.

[47] Arnold M. Weissler, William S. Harris, and Clyde D. Schoenfeld. Systolic time intervals in heart failure in man. *Circulation*, 37(2):149–159, 1968.

[48] Ziheng Xi, Jiayi Yu, Yitao Wang, Yanbo Duan, Jianjiang Feng, and Jie Zhou. EgoEMG: A multimodal egocentric dataset with bilateral EMG and vision for hand pose estimation. *arXiv preprint arXiv:2605.05712*, 2026.

[49] Yan Xia, Xiang Chen, Nishant Ravikumar, Christopher Kelly, Rahman Attar, Nay Aung, Stefan Neubauer, Steffen E. Petersen, and Alejandro F. Frangi. Automatic 3D+t four-chamber CMR quantification of the UK Biobank: integrating imaging and non-imaging data priors at scale. *Medical Image Analysis*, 80:102498, 2022.

[50] Guoqi Yu, Juncheng Wang, Chen Yang, Jing Qin, Angelica I Aviles-Rivero, and Shujun Wang. Decentralized attention fails centralized signals: Rethinking transformers for medical time series. In *The Fourteenth International Conference on Learning Representations*, 2026.

[51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023.

[52] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024.

A Appendix

A.1 Dataset and Code Access

The heart mesh and ECG data can be accessed in the UK Biobank after applying for permission at this URL https://www.ukbiobank.ac.uk/. Code can be accessed at this URL https://github.com/Chrisa142857/time2track

A.2 Implementation Details and Hyperparameters

Time2Track is implemented in PyTorch with HuggingFace `accelerate`. Unless stated otherwise, all baselines share the diffusion backbone (Spatio-Temporal Transformer, $L_{\text{dit}} = 8$ layers, latent width $D = 256$, $N_{\text{patch}}{+}C \cdot L = 48$ conditioning tokens) and only the electrophysiology encoder is varied, so observed differences are attributable to the conditioning representation. Training and inference settings are summarized in Table 4.

α-gating schedule and wavelet fine-tuning. The delay-branch gate α is a single learnable scalar initialized to 0, so at step 0 the model reduces exactly to the PhysioWave morphology baseline. α is then updated freely by AdamW (no manual schedule) and grows as the delay tokens become useful. The wavelet front-end is initialized from the publicly released PhysioWave ECG checkpoint and not frozen: every parameter (template mixing weights, learnable kernels, and the soft-argmax temperatures $\{\beta_\ell^{\text{raw}}\}$) is updated end-to-end with the diffusion loss.

Dataset preprocessing. Resting 12-lead ECG (500Hz) is cropped to a single 600-sample window aligned to the cardiac MRI cycle and z-normalized per channel. Cardiac meshes have $N{=}388$ control points and are sampled to $F{=}50$ frames per cycle. Subjects with missing frames are filtered, yielding $n{=}40{,}577$ ECG/mesh pairs split 80/20 train/val with seed 42.

Table 4: Training and model hyperparameters for Time2Track on UKB ECG→4D heart mesh.

Optimization	
Optimizer / (β_1, β_2) / ϵ	AdamW / $(0.9, 0.999)$ / 10^{-8}
Weight decay	1×10^{-2}
Learning rate / schedule / warmup	1×10^{-5} / cosine / 200 steps
Per-GPU batch size / grad-accum / effective batch	20 / 4 / 320 (4×GPUs)
Epochs / max grad norm / mixed precision	200 / 1.0 / bf16
Dropout / cond. drop rate (CFG train)	0.1 / 0.1
Gradient checkpointing / xFormers attn.	enabled / enabled
Diffusion	
Backbone / layers / latent dim D	ST-Transformer / 8 / 256
Inference sampler / steps K_{infer} / CFG scale	DDIM / 50 / 2.0
Trajectory frames F / control points N	50 / 388
Velocity loss weight λ_{vel}	1.0
Morphology branch	
Levels L / wavelet kernel size	3 / 24
Wavelet template $\{\psi_m\}$ [8]	db4, db6, sym4, coif2
Patch length P / morph. embed dim D'	64 / 384
Depth / heads	8 / 12
Init. weights / fine-tuning	[8] / full fine-tune
Delay branch	
Reference lead r	lead II
Search window τ_{max} (samples / ms @500Hz)	$\pm$100ms
Soft-argmax temperature init. β_ℓ^{raw}	softplus(2.0)
Fourier lag bands K / freqs. ω_k	8 / log-spaced, $2^k\pi/\tau_{\text{max}}$
Delay transformer heads	4
Gate α initialization	0
Hardware / runtime	
GPUs / wall-clock time	4×NVIDIA H100 NVL / ≈25h to convergence

A.3 Ablation Studies

We isolate the contribution of the delay branch and of the FFT-CC delay extractor by holding the diffusion backbone and training schedule of Sec. A.2 fixed and varying only the electrophysiology encoder. Table 5 reports whole-heart 3D vertex metrics on the UKB validation set. The $|\text{APCR} - 1|$ and $|\text{APDR}-1|$ columns directly measure inter-subject diversity collapse. All numbers are reproduced from the corresponding rows of Table 1 for direct comparison.

Effect of the delay branch. Removing the delay branch costs roughly $4\times$ on $|\text{APDR} - 1|$ (0.019→0.079), confirming that delay tokens are the primary lever against collapsinng to the average.

Effect of FFT-CC. Replacing FFT-CC with a single T-peak or QRS-peak delay (Delay-T / Delay-QRS) shows only minor improvement of the inter-subject diversity and loses local fidelity (PAR drops from 0.896 to 0.772 for Delay-T).

Skip the FFT-CC and use a directly regressed scalar lag per (channel, scale) pair (→ *raw scalar lag*) leaves worse point-wise error and inflates inter-subject diversity (|APCR-1| 0.081 vs. 0.034) and shrinks |APDR-1| by 0.195. The differentiable cross-correlation provides a sharper, less noisy lag signal than direct regression.

Effect of the lag embedding. Both removing the Fourier-feature embedding (feeding the scalar lag through an MLP) and removing positional embeddings $\mathbf{u}, \mathbf{v}$ on the delay tokens have a worse RV and APDR than the full Time2Track, confirming that the NeRF-style $\sin / \cos$ bands and channel/scale positions both carry signal that the backbone cannot recover from a scalar.

Table 5: Component ablations on UKB whole-heart vertex trajectories. "Ours" is the full Time2Track implemented with hyperparameters listed in Table 4.

Variant	RV ↓	RMSE ↓	PAR → 1	APCR → 1	APDR → 1
Delay branch					
w/o delay branch	0.634	0.172	0.868	0.956	0.826
FFT-CC					
→ Delay-T	0.619	0.158	0.772	1.042	0.743
→ Delay-QRS	0.589	0.166	0.894	1.042	0.907
→ raw scalar lag	0.615	0.178	0.887	1.081	0.786
Fourier-feature Lag embedding					
→ scalar + MLP	0.597	0.179	**0.896**	1.084	0.923
w/o $\mathbf{u}, \mathbf{v}$	0.595	0.160	0.858	1.042	0.790
Morphology branch					
Patch length (P=4)	0.588	**0.142**	0.275	2.398	0.139
Patch length (P=8)	0.512	0.144	0.726	2.275	0.472
Patch length (P=16)	0.545	0.145	0.276	2.391	0.144
Ours (full)	**0.637**	0.151	**0.896**	**0.966**	**0.981**

Effect of patch length for the morphology branch. Varying the patch length (4 / 8 / 16) *collapses* amplitude and inter-subject diverisity (PAR ≤ 0.726, APDR ≤ 0.472, APCR ≥ 2.275). The full model's bank of {db4, db6, sym4, coif2} with 64 patches strikes the right multi-resolution coverage; both under- and over-parameterized banks lose the electromechanical phase information that the delay branch is meant to inject.

A.4 Encoder Properties

Table 6 lists the criterion for including every encoder, and its column of time lag covers no lag, shift invariance, classical T-peak and QRS delays, and the learned lag. Compared with the diversity in Table 1, the parameter count does not order the encoders, whereas the handling of time lag does: $|\text{APDR} - 1|$ is 0.151 and 0.170 without lag, 0.079 with wavelet scales, and 0.019 with the learned lag.

Table 6: Properties of the compared electrophysiology encoders.

Encoder	Architecture	Time lag	Why included	Params
CNN	1D conv on raw time	none	vanilla	0.21 M
TeCh	attention over time patches and leads	none	SOTA MedTS Transformer	8.16 M
PhysioWave	attention over (scale, time), wavelets	none	SOTA pretrained ECG	14.35 M
ShiftInvariant	shift invariant net, no attention	invariant	SOTA shift invariance	0.15 M
Delay-T	TeCh plus external delay matrix	classical T peak	classical delay	9.21 M
Delay-QRS	TeCh plus external delay matrix	classical QRS	classical delay	9.21 M
Ours	attention, learnable wavelets, delay	learned, gated	proposed	15.41 M

A.5 Error Bars of the Diversity Metrics

APCR and APDR are pairwise statistics over the diversity sample of 128 subjects, so their uncertainty is estimated by a paired subject-level bootstrap with 10,000 resamples, in which the numerator and the denominator of each ratio are recomputed on the same resample. The $\pm$ of $|\text{APCR} - 1|$ and $|\text{APDR} - 1|$ in Table 1 is the SD of $|\text{APCR} - 1|$ and $|\text{APDR} - 1|$ across the resamples, which ranges from 0.037 to 0.078 and from 0.006 to 0.012, respectively.

A.6 Robustness to Electrode Noise, Missing Leads and τ_{max}

Table 7 reports two tests on the diseased subgroup, per-lead Gaussian noise added at inference with σ scaled from the power of each lead to reach the target SNR, and lead channels set to zero at inference. Dropping the reference lead II leaves fidelity intact but raises $|\text{APCR} - 1|$, and the model still generalizes when three chest leads are dropped. Table 8 shows that a $4\times$ change of τ_{max} at inference keeps RV and RMSE within their bootstrap intervals and changes PAR, $|\text{APCR} - 1|$ and $|\text{APDR} - 1|$ by at most 0.011.

Table 7: Electrode noise and lead dropouts at inference on the diseased subgroup.

Input	RV↑	RMSE↓	PAR $\to 1$	$\lvert\text{APCR} - 1\rvert \downarrow$	$\lvert\text{APDR} - 1\rvert \downarrow$
clean	0.576	0.179	0.892	0.027	0.014
noise, 20 dB SNR	0.576	0.179	0.891	0.027	0.014
noise, 10 dB SNR	0.572	0.180	0.882	0.020	0.012
noise, 5 dB SNR	0.562	0.183	0.871	0.002	0.010
drop V2 (single chest)	0.579	0.178	0.888	0.024	0.017
drop II (reference)	0.575	0.183	0.958	0.108	0.011
drop V1, V2, V3 (three chest)	0.581	0.177	0.891	0.050	0.021

Table 8: Sensitivity to τ_{max} at inference, default in bold (95% bootstrap intervals for RV and RMSE).

τ_{max}	RV↑	RMSE↓	PAR $\to 1$	$\lvert\text{APCR} - 1\rvert \downarrow$	$\lvert\text{APDR} - 1\rvert \downarrow$
25 ($\pm$50 ms)	0.570 [0.557, 0.585]	0.181 [0.177, 0.185]	0.898	0.052	0.011
50 ($\pm$100 ms)	0.568 [0.555, 0.581]	0.181 [0.177, 0.186]	0.902	0.050	0.011
75 ($\pm$150 ms)	0.567 [0.552, 0.581]	0.181 [0.177, 0.186]	0.901	0.041	0.012
100 ($\pm$200 ms)	0.565 [0.552, 0.579]	0.181 [0.177, 0.186]	0.903	0.046	0.012

A.7 Reference Lead

Lead II is the standard reference lead in clinical electrocardiography owing to its high SNR [19]. Table 9 shows that retraining Time2Track with aVR or aVL as the reference lowers RV and raises RMSE, $|\text{APCR} - 1|$ and $|\text{APDR} - 1|$. In a preliminary ES timing classification without a reference lead, aggregating the full delay map by its mean or median reaches an AUC of 0.590 or 0.563, against 0.597 for the lead II delay. The median inter-lead T-peak delay in UKB is 8 ms (interquartile range 2 to 30 ms), so the 2 ms sampling interval at 500 Hz does not limit its quantification.

Table 9: Sensitivity to the reference lead.

Reference lead	RV↑	RMSE↓	PAR $\to 1$	$\lvert\text{APCR} - 1\rvert \downarrow$	$\lvert\text{APDR} - 1\rvert \downarrow$
II (Table 1)	0.637	0.151	0.896	0.034	0.019
aVR	0.590	0.181	0.948	0.041	0.022
aVL	0.592	0.181	0.934	0.074	0.024

A.8 Delay Tokens for sEMG Decoders

Delay descriptor. The sEMG delay branch keeps the FFT-CC of Eq. 6, the soft-argmax with a learnable temperature and the per-channel delay tokens of Sec. 3.1, with two changes for sEMG. First, sEMG channels record asynchronous motor-unit interference patterns rather than delayed copies of one propagating wave, so the branch cross-correlates the activation envelope of each channel instead of its wavelet detail bands. Specifically, the envelope of channel c is the rectified signal $|x_c|$ after mean removal and an FFT low-pass filter at 15 Hz, and the soft-argmax over $|\tau| \le \tau_{\max}$, with β^{raw} initialized at 50, yields one lag δ_c per channel relative to the first of the 16 channels as the reference r. Second, each δ_c is embedded by a two-layer MLP ($1 \to 1024 \to 256$ with GELU) instead of the Fourier feature lag embedding and summed with a learnable channel embedding $\mathbf{u}_c$, and one Transformer block (4 heads, MLP ratio 4, dropout 0.1) mixes the 16 channel tokens into delay tokens

of width 256. The 2 kHz sEMG window is split into K chunks, and every chunk yields its own 16 delay tokens.

FiLM injection into vemg2pose. vemg2pose [34] decodes hand pose with a time-depth separable (TDS) convolutional featurizer followed by an LSTM and has no token sequence. Therefore, we inject the delay tokens by feature-wise linear modulation (FiLM) [27] of the 64 featurizer channels $\mathbf{h}_t$, namely $\mathbf{h}'_t = \mathbf{h}_t \odot (1 + \mathbf{g}_t) + \mathbf{b}_t$, where a linear head maps the channel average of the delay tokens of each chunk to a gain $\mathbf{g}$ and a shift $\mathbf{b}$, and linear interpolation over the chunks gives $(\mathbf{g}_t, \mathbf{b}_t)$ at every output frame t. The branch reads $K = 2$ chunks of 1,024 samples (about one median gesture cycle), which end at the start and at the middle of the 2,048-sample output window, with $\tau_{\max} = 200$ samples ($\pm$100 ms). The weights and bias of the linear head are initialized to zero as in the zero convolution of ControlNet [51], so the modulated decoder equals the released vemg2pose at initialization while the head still receives gradient. This zero convolution replaces the zero-initialized scalar gate α of the cardiac model, which remained near zero in our sEMG runs because the output head receives no gradient while $\alpha = 0$.

Cross-attention injection into EMGFormer. EMGFormer [48] decodes hand pose with a TDS featurizer followed by a Transformer encoder (3 layers, width 256, 4 heads, rotary position embedding) over frame tokens. Since the encoder already operates on a token sequence, the delay tokens enter without channel averaging. The branch reads $K = 3$ chunks of 2,048 samples spread evenly over the 7,280 output samples of the 7,790-sample input window, with $\tau_{\max} = 800$ samples ($\pm$400 ms), which yields $3 \times 16 = 48$ delay tokens $\mathbf{U}$. After the last encoder layer, every frame token $\mathbf{z}_t$ queries the 48 delay tokens through a 4-head cross-attention,

$$\mathbf{z}'_t = \mathbf{z}_t + \mathbf{W}_o \,\mathrm{MHA}\big(\mathbf{z}_t, \mathrm{LN}(\mathbf{U}), \mathrm{LN}(\mathbf{U})\big), \tag{12}$$

where LN denotes layer normalization and the zero-initialized linear layer $\mathbf{W}_o$ serves as the zero convolution, so the decoder equals the released EMGFormer at initialization. In contrast to FiLM, the cross-attention assigns every output frame its own weighting of the 48 delay tokens.

Training and evaluation. Each treated decoder and its control without delay tokens are fine-tuned from the released checkpoint with the same training windows, loss and budget, so the two arms differ only in the delay branch. Both decoders use the official emg2pose objective, namely the joint angle MAE plus 0.01 times the fingertip distance. vemg2pose is fine-tuned for 6,000 steps by Adam with a learning rate of 10^{-4}, a batch of 64 windows of 3,838 samples (the 1,790-sample left context and the 2,048-sample output window), random electrode rotation by up to one channel and gradient clipping at 1.0, and the final weights are evaluated. EMGFormer follows its official recipe, namely AdamW with a learning rate of 10^{-4}, cosine decay to 5×10^{-5}, weight decay 0.01, an effective batch of 650, bf16 and gradient clipping at 1.0, for 60 epochs (1,560 steps), and the weights with the lowest validation MAE are evaluated. Every training and test window of a session begins at the first usable gesture cycle of that session, detected from peaks of the low-pass filtered joint speed of the ground truth. We score 49 frames over the 2,048 samples after the cycle onset by RMSE, PAR, APCR and APDR, the metrics of the cardiac comparison in Table 1, with APCR and APDR computed across the 178 test users, on 6,151 test sessions for vemg2pose and 6,149 for EMGFormer. Every arm is trained with the seeds 42, 43 and 44.

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: See Abstract and Introduction.

   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: See the section Conclusions and Limitations.

   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:[N/A]

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: See Method and Appendix sections.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: See Appendix section.

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: See Experimental Results and Appendix sections.

   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: See Fig. 1 for significance and error bars present in the Experimental Results section.

   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: See Appendix section.

   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: Our research follows the NeurIPS Code of Ethics.

   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: [N/A]

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: [N/A]

    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: UK Biobank application number will be included in the Acknowledgments.

    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: [N/A]

    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: [N/A]

    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: [N/A]

    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: [N/A]

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.