TOWARD WORLD MODELING OF PHYSIOLOGICAL SIGNAL LEVERAGING LATENT DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Physiological time series signals reflect complex, multi-scale dynamical processes of the human body. Existing modeling studies focus on static tasks such as classification, event forecasting, or short-horizon next step prediction, while long-horizon signal-level forecasting and predictive nature of physiological signals remain underexplored. We introduce **NormWear-2**, a world model that encodes both multivariate physiological signals and clinical intervention variables into a shared latent space and models their joint temporal evolution as a dynamical system. Our approach combines inference from prior pre-trained knowledge (*intuition*) with instant non-parametric latent state transition adaptation (*insight*), enabling coherent forecasting across multiple temporal scales, conditioned on heterogeneous clinical interventions. During the pretraining phase, we find that chaos-theoretic balancing of dynamical regime diversity yields more robust representations, with a smaller balanced corpus outperforming one twice its size and capturing bifurcation regimes. We evaluate the world model performance across diverse real-world physiological datasets spanning heterogeneous temporal resolutions and intervention regimes, covering daily life, point-of-care, and clinical settings, including fitness planning, hemodialysis, diabetes management, and surgical monitoring. These evaluation datasets comprise records from 8,026 subjects, spanning study durations from 3.2 hours for high-resolution signal data to 2.3 years for longitudinal clinical biomarker tracking. NormWear-2 achieves the best overall forecasting performance across time, frequency, and latent representation domains, with significant improvements over state-of-the-art time series foundation models, while maintaining competitive downstream representation quality, providing a step toward general-purpose world models for physiological signals.

1 INTRODUCTION

Physiological signals provide a continuous and non-invasive window into the internal dynamics of the human body. Modalities such as electroencephalography (EEG), electrocardiography (ECG), and photoplethysmography (PPG) encode rich temporal patterns spanning multiple scales, from milliseconds to hours. These signals are inherently generated by complex dynamical systems, yet most machine learning approaches treat them as static inputs for downstream tasks such as classification, regression, or anomaly detection (Pillai et al., 2024; Lee et al., 2026; Luo et al., 2026). As a result, they lack the ability to reason about future trajectories, simulate alternative scenarios, or adapt to changing physiological states.

More broadly, real-world multivariate time series often arise from systems exhibiting diverse dynamical behaviors, ranging from quasi-periodic and limit-cycle patterns to weakly or strongly chaotic regimes (Strogatz, 2024; Wu et al., 2021; Tan et al., 2025; Lai et al., 2026). Differences in observed time series frequently reflect variations in predictability, sensitivity to initial conditions, and structural complexity rather than fundamentally distinct generative mechanisms. Capturing such nonlinear dynamics is critical for learning representations that generalize across systems and domains. Recent advances in representation learning have improved performance in general time-series modeling (Ansari et al., 2025; Liu et al., 2025; Lai et al., 2026; Gilpin, 2021). However, these works did not explicitly investigate in the distribution of underlying dynamical regimes while pretrain the base model. This can not only diminish our understanding toward the behaviors of model learning and inference, but also limit the robustness and transferability of learned representations. Further-

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

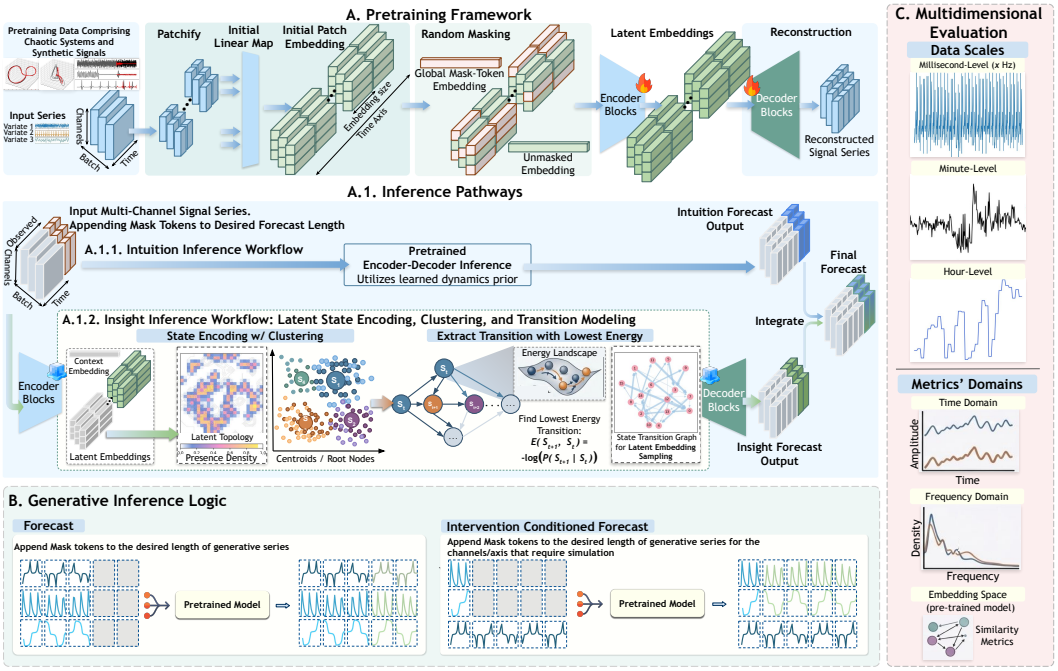

Figure 1: **Methodology.** (A) Overview of the modeling workflow from the input signals to pretraining and forecasting output. (A.1.) Proposed intuition-insight inference pathways. (B) Demonstration of the generative prediction logic after the standard mask-and-reconstruction pretraining. (C) Multidimensional evaluation across multiple temporal resolution and performance metrics.

more, the feasibility of these approaches on physiological signals and digital healthcare applications involving intervention variables remain underexplored.

To address these challenges, we propose NormWear-2, a framework for latent world modeling of physiological signals as briefly demonstrated in Figure 1. The model encodes multimodal time series into a shared latent space where temporal evolution is modeled as a dynamical system, combining two complementary mechanisms: (1) *intuition*, capturing prior knowledge of physiological dynamics learned from large-scale pretraining, and (2) *insight*, refining latent state transitions through instant non-parametric unsupervised adaptation. This dual mechanism enables coherent multi-step forecasting and generative modeling while adapting to new observations and contexts. During the modeling exploration, we also discover that the effectiveness of latent world modeling depends on the diversity of dynamical regimes present in pretraining data. Leveraging chaos-inspired metrics such as Lyapunov exponents, detrended fluctuation analysis, and persistent entropy (Wolf et al., 1985; Hu et al., 2001; Atienza et al., 2019), we quantify dynamical properties of time series and construct pretraining corpora with balanced coverage across different dynamical behaviors. Empirically, models pretrained on dynamically balanced datasets produce more robust and transferable representations, supporting downstream predictive and generative tasks. Finally, we introduce a multidimensional evaluation framework that assesses forecasting quality across temporal, frequency, and latent representation domains. Experiments across a diverse range of real-world physiological datasets demonstrate the effectiveness of NormWear-2 for multi-scale forecasting, indicating latent dynamical modeling as a promising paradigm for predictive physiological signal modeling.

2 Related Work

In physiological domains, most approaches, ranging from task-specific architectures (Foumani et al., 2024; Wang et al., 2025; McKeen et al., 2025) to pretrained representations (Luo et al., 2026), emphasize supervised or self-supervised representation learning for downstream tasks, rather than explicit modeling of temporal dynamics for forecasting or simulation. This limits their ability to support predictive reasoning under interventions. Dynamical systems approaches, including state-space and latent variable models, provide a principled framework for modeling temporal evolution (Wang et al., 2024). More recently, world models and latent predictive learning have emphasized learning structured representations for forecasting (Maes et al., 2026; Nam et al., 2026). From a modeling

Table 1: Datasets for evaluation. The time duration in this table referring to the average time spanning of the data for each subject. Additional information of each dataset is provided in Appendix A.

Dataset	# Subjects	Time Duration	Scale & Scope	Physiological Signals	Intervention Variables
VitalDB (Lee et al., 2022)	6,388	3.2 Hours	Millisecond, Wearable and Medical Device Sensing	ECG, PPG, EEG, Respiration	Remifentanil Target Concentration, Fraction of Inspired Oxygen, Fresh Gas Flow Rate (L/min), Inspiratory Time (s), Respiratory Rate, Tidal Volume (mL)
PMData (Thambawita et al., 2020)	16	5 Months	Minute, Sport Wristband Sensing	Heart Rate, Steps, Distance, Calories	Lifestyle Sports, e.g. Running, Soccer, Strength, etc.
CGMacros (Gutierrez-Osuna et al., 2025)	45	10.9 days	Minute, Biofluidic Sensing	Glucose, Heart Rate, Physical Motion	Food Nutrition Composition
Shanghai Diabetes (Zhao et al., 2023)	125	10.7 days	Quarter Hour, Biofluidic Sensing	Glucose, Heart Rate	Insulin, Hypoglycemic Agents
KidneyDialysis (Luo et al., 2024)	1,452	2.3 Years	Hour, Medical Device Sensing	Heart Rate, Blood Pressures, Body Temperature	Rates of Blood Flow and Dialysate Flow, Dialysate Temperature, Ultrafiltration Rate

perspective, existing world modeling approaches can be broadly categorized into generative-based methods (Ansari et al., 2025; Lai et al., 2026; Liu et al., 2025) and joint-embedding predictive architectures (JEPA) (Maes et al., 2026). Generative approaches explicitly model the conditional distribution of future trajectories. Recent time series foundation models (Das et al., 2024; Rasul et al., 2023; Woo et al., 2024; Wu et al., 2021) provide a strong backbone for this paradigm, while model inference behavior across varied temporal resolutions, and the feasibility of applying such models to physiological signal forecasting in scenarios involving intervention variables, also referred to as world modeling or physiological simulation, remain largely underexplored. In contrast, JEPA-based approaches (Fox et al., 2025) primarily focus on learning predictive representations and have been explored in physiological domains mainly for improving downstream inference tasks, rather than signal level forecasting accuracy. As a result, this work fills these research gaps and proposes a well-rounded, end-to-end methodology for world modeling in physiology and digital health.

3 METHOD

3.1 PROBLEM SETUP

World modeling (Maes et al., 2026; Hu & Shu, 2023) refers to a model that maintains an internal understanding of the world and is able to forecast future trajectories given current and past observations. In the field of physiological signal modeling, such a paradigm remains under-explored, as discussed in the introduction and related work. In the context of digital healthcare, we define the world model as a dynamical system, which is commonly formulated as $\mathbf{x}_{t+1} = f_\theta(\mathbf{x}_{\leq t}, \mathbf{u}_{\leq t})$, where $\mathbf{x}_t$ denotes the system state and $\mathbf{u}_t$ represents the action or intervention at time t. In the context of physiological signals, we instantiate this formulation as a *multivariate time series system*, where both physiological measurements and interventions evolve over time. This naturally leads to a conditional forecasting formulation: $p_\theta(\mathbf{x}_{t+1:t+H} \mid \mathbf{x}_{\leq t}, \mathbf{u}_{\leq t+H})$, where H denotes the prediction horizon. Under this view, world modeling in healthcare becomes a channel-conditioned multivariate time series forecasting problem, because in many clinical scenarios, interventions are inherently temporal and can be directly aligned with physiological signals. For instance, in surgical settings, machine-controlled parameters such as respiration and anesthesia delivery are continuous time series that evolve synchronously with physiological variables. Similar patterns arise in dialysis and other critical care scenarios. For event-based interventions, such as insulin administration or meal intake in diabetes management, we convert them into step-function representations series, where non-zero values span the duration of the event. More abstract lifestyle factors, such as physical activity, can be encoded via a pretrained language model (Alsentzer et al., 2019) into a finite semantic space during data preprocessing, and then transformed in the same way as the event-based interventions. These strategies enable a unified temporal representation of heterogeneous interventions across diverse healthcare scenarios.

3.2 DATASETS

To study world modeling of physiological signals, we require datasets that are multivariate, longitudinal, and include explicit action/intervention variables. We leverage five real-world physiological datasets for core method evaluation: VitalDB (Lee et al., 2022), a perioperative monitoring dataset with physiological signals and anesthesia machine parameters for surgical patients; PMData (Thambawita et al., 2020), a wearable health dataset with heart rate, activity, sleep, and exercise records from daily life monitoring; Shanghai Diabetes (Zhao et al., 2023), a real-world diabetes dataset with glucose records, medications, insulin, and dietary information; CGMacros (Gutierrez-Osuna et al.,

2025), a diabetes dataset with glucose monitoring, meal records, and activity information for personalized nutrition analysis; and a clinical kidney dialysis (KidneyDialysis) dataset (Luo et al., 2024), a hemodialysis dataset with physiological signals and dialysis machine parameters during treatment sessions. These datasets cover representative healthcare scenarios including clinical monitoring, daily health management, point of care treatment, and clinical treatment, providing diverse physiological observations and intervention variables that make them well suited for evaluating healthcare world models. For downstream tasks, we follow the same evaluation datasets as in Luo et al. (2026) and adopt the same evaluation setup. For studying data-balance-aware pretraining, we primarily use datasets from (Lai et al., 2026; Luo et al., 2026). As this preliminary exploration study is closely related to dynamical system modeling, we additionally include benchmark datasets from prior works for validation, including Wu et al. (2021); Tan et al. (2025); Gilpin (2021). Detailed datasets information are reported in Appendix A.

3.3 Chaos-Theoretic Metrics for Data-Balance-Aware Pretraining

We leverage a set of metrics from nonlinear dynamical systems and chaos theory to characterize and quantify different types of chaotic processes present in a collection of observed time series. These metrics include the detrended fluctuation analysis (DFA) exponent (Hu et al., 2001), the Lyapunov exponent (LE) (Wolf et al., 1985; Kantz & Schreiber, 2003), and persistent entropy (PE) computed on zero- and one-dimensional homology (Atienza et al., 2019; 2020). Respectively, these metrics assess long-range autocorrelation, sensitivity to initial conditions, and the connectivity and loop complexity of the transformed topological structure of a time series. Based on these chaos metrics, different dynamical system types can be identified through a deterministic pipeline that computes the metrics, applies unsupervised clustering, and assigns cluster labels using fixed, literature-established thresholds (Hu et al., 2001; Kantz & Schreiber, 2003; Atienza et al., 2019). To examine dataset balance with respect to chaotic behavior, we perform K-means clustering on the computed metrics for each time series sample. The optimal number of clusters is selected using the elbow method. Details of the clustering procedure and the interpretation of each cluster in terms of underlying dynamical systems are provided in Appendix C.4 and Appendix E.1. As shown in Figure 2 (panel A), the pretrained multivariate time-series benchmark datasets from Lai et al. (2026) and Luo et al. (2026) are each dominated by a single type of chaotic pattern. In contrast, when these benchmarks are aggregated, the resulting dataset exhibits a more homogeneous distribution of chaotic types.

3.4 Pretraining and Generative Inference

Model backbone. We use the channel-aware mechanism for multivariate signal modeling proposed by Luo et al. (2026). Detailed complexity analysis of multivariate time series modeling approaches (Luo et al., 2026; Lai et al., 2026; Liu et al., 2025) are presented in Appendix B.5. Figure 1.A., with the same encoding backbone as proposed by Luo et al. (2026), but optimized logic for initial time series patch embedding and the lightweight decoding block to better adapt the scenario for generative based multivariate time series modeling purpose.

Training and Inference. The backbone model is pretrained on the aggregated pretraining data benchmark as discussed in Section 3.3, in a masking and reconstructing manner. After the input multivariate time series being patchified, the patches are randomly replaced with a trainable unified [MASK] token representation with a fixed probability threshold pre-defined following guidance from Huang et al. (2022). The masks applied on the input are independently sampled for each channel in the multivariate input, thus, varied masking combination are expected to be covered as more pretraining iterations progresses. During inference time, we focus on two types of generative tasks in this study: forecasting and simulation. For forecasting, the model predicts the future time series data given the past time series data. On the other hand, simulation task involves completing the unobserved channel conditioned on one or more given or observed channel. We refer this task as simulation because it naturally align with varied application scenarios such as health intervention and battery testing where we may have one or more controlled variables represented as separate input time series channels. The overview is presented in Figure 1 (panel E).

3.5 Dynamical State Transition Modeling in Latent Space

After pretraining, the dynamical state transition modeling is performed during inference time. Given an observed context, the input is first encoded into latent representations via the pretrained encoder, which are then grouped into clusters to form discrete latent states. The optimal number of clusters

is determined using the elbow rule, and empirically we observe that it scales approximately with the logarithm of the total number of patches extracted from the observed context. To capture temporal dynamics, we estimate state transitions directly from consecutive patch pairs. Let $s_t \in \mathbb{R}^{emb_size}$ denote the latent cluster assignment at time step t where emb_size representing the embedding size of the backbone model. The empirical transition probability is

$$P(s_{t+1} = j \mid s_t = i) = \frac{\sum_t \mathbf{I}[s_t=i,\, s_{t+1}=j]}{\sum_t \mathbf{I}[s_t=i]}, \qquad s'_{t+1} \sim \sum_j P(s_{t+1} = j \mid s_t = i)\mathcal{N}(\mu_j, \sigma_j^2), \qquad \mu_j, \sigma_j^2 \in \mathbb{R}^{emb_size} \quad (1)$$

where $\mathbf{I}[\cdot]$ is the indicator function, and μ_j and σ_j denote the centroid and standard deviation of cluster j. Thus, forecasting is performed by first sampling the next latent state from the transition matrix and then drawing the latent representation from the corresponding Gaussian component. When intervention or action variables are available, the state transition becomes action-conditioned. Specifically, the transition probability from state i to state j is no longer determined solely by the current state, but also by the applied action a_t. The marginal transition can be expressed as

$$P(s_{t+1} = j \mid s_t = i) = \sum_{a \in A} P(s_{t+1} = j \mid a, s_t = i)\, P(a \mid s_t = i), \quad s'_{t+1} \sim \sum_j P(s_{t+1} = j \mid a_t, s_t = i)\mathcal{N}(\mu_j, \sigma_j^2) \quad (2)$$

where the first term describes the marginalization over possible actions A, and the second term represents the action-conditioned forecasting distribution. This formulation is equivalent to the decomposition of the expected transition output presented in Equation 1. When action variables are introduced as additional multivariate input time-series channels, inference becomes more straightforward. At each time step, the latent state representation is concatenated with the corresponding action vector, forming an action-aware state embedding. The transition decomposition can then be directly obtained through straightforward Euclidean distance based neighborhood search over these joint state-action representations, and state transitions are subsequently restricted to the retrieved neighboring states, ensuring that forecasting remains consistent with the observed intervention dynamics. This transition model can alternatively be interpreted in an energy-based formulation, where an energy function is defined as $E(s_t, s_{t+1}) = -\log P(s_{t+1}|s_t)$. Under this view, forecasting corresponds to sampling low-energy transitions, which is equivalent to maximizing the likelihood in the corresponding Markov graphical model. After generating latent representations for the desired number of future time steps, all patches are passed through the pretrained decoder to reconstruct the output time series. We refer to the direct output from the pretrained model as "intuition", as it depends entirely on the pretrained backbone, while the proposed dynamical transition modeling stage is termed "insight", since it incorporates information from the observed context during inference.

3.6 Multidimensional Evaluation

Beyond evaluating performance across multiple temporal scales, we propose a multidimensional evaluation protocol that captures diverse aspects of forecasting quality. Rather than relying solely on step-wise deviation from ground truth, our evaluation framework incorporates complementary metrics that assess structural, spectral, and representation-level alignment. Specifically, we consider three categories of metrics. First, for point-wise accuracy, we adopt Mean Absolute Error (MAE) to measure the average deviation between predicted and ground-truth sequences. Second, to account for temporal and morphological consistency, we employ Dynamic Time Warping (DTW), implemented via the differentiable SoftDTW formulation, which allows flexible alignment between sequences with temporal distortions. Third, to evaluate frequency-domain characteristics, we compute the cosine similarity (FreqCosSim) and Euclidean distance (FreqEucl) between the Fast Fourier Transform representations of predicted and ground-truth signals, capturing discrepancies in spectral components. Furthermore, to assess high-level semantic consistency, we leverage a pretrained encoder to extract latent embeddings from both predicted and ground-truth sequences. We then compute cosine similarity (LatentCosSim) and Euclidean distance (LatentEucl) in the latent space, providing a measure of abstract representation alignment beyond observable signal space. A final score is also keep track during the experiments to better compare the overall performance of different approaches. Each test sample is Z-normalized on the observed context, and the ground truth of unobserved part, which is to be predicted by the model, is also z-normalized based the mean and standard-deviation from the observed context. The scores are normalized to make sure all the metrics live on consistent numerical scales, thus, we have final score computed as:

$$\frac{1}{6} \cdot \left(MAE + \frac{Soft_DTW}{pred_length} + (1 - FreqCosSim) + \frac{FreqEucl}{0.5 \cdot pred_length} + (1 - LatentCosSim) + \frac{LatentEucl}{embed_size} \right) \quad (3)$$

Where $pred_length$ depends on the application scenario as specified in Table 4, and scale it by 0.5 representing the max number of unique frequency components. Together, these metrics offer a comprehensive evaluation framework that holistically reflects forecasting performance across temporal alignment, spectral fidelity, and semantic representation.

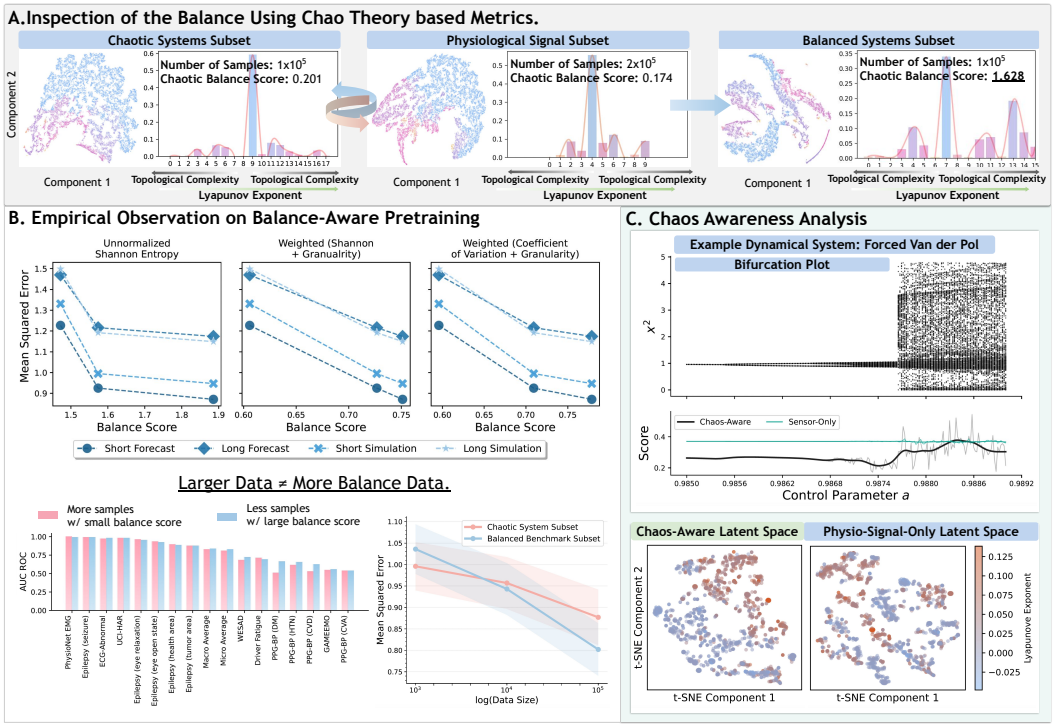

Figure 2: (A) Inspection of the balance using chao theory based metrics. (B) Balance-aware behavior: We show performance on generative tasks across models pre-trained from datasets with varied balance score (metrics details in Appendix C). (C) Chaos-Aware analysis.

4 Experiment and Result

4.1 Preliminary Exploration on Chaos-Balance-Aware Pretraining

To investigate the effect of chaos-balance-aware pretraining, we evaluate models on both generative and downstream tasks, including forecasting, simulation, classification, and regression. Generation quality is measured by MAE/MSE at this stage, while downstream performance is assessed using task-specific metrics following Luo et al. (2026). To isolate the role of dataset balance, we construct controlled pretraining subsets with identical data size (10^5 samples) but different balance scores, where balance is quantified using weighted Shannon entropy and granularity-based diversity measures over clustered dynamical systems (Appendix C.2). As shown in Figure 2 panel B, models pretrained on more balanced subsets consistently achieve lower generative error across varied evaluation settings. We then compare a smaller but more balanced corpus (10^5 samples, balance score 0.73) against a larger but less balanced corpus (2×10^5 samples, balance score 0.60), while keeping architecture and optimization settings fixed. As shown in the lower half of panel B in Figure 2, despite the reduced data size, the more balanced pretraining set yields better performance across multiple scenarios. We further observe that data scaling and balance interact synergistically: when increasing pretraining data size under fixed model capacity, more balanced datasets exhibit faster reductions in test-time generative error. Together, these preliminary findings suggest that chaos-balance-aware curation tends to be the one of the core aspects that can be intuitively identified that provides the model a decent generative forecasting quality. We provide supplementary empirical details in Appendix D.8.

Finally, we evaluate chaos awareness using the Forced Van der Pol as the paradigm system (Figure 2C). Its bifurcation diagram shows that in stable regions the system converges to one or two local optima, while increasing the forcing parameter drives the system into chaos; for example, when $a \geq 0.9875$, the optima become highly dispersed, indicating strong chaoticity. We compare forecasting performance between a chaos-balanced pre-trained model and a sensing-only model by conditioning on the first 2048 of 4096 sampled steps and predicting the remaining steps. The chaos-balanced model performs well in stable regions and degrades mainly after the onset of chaos, while still maintaining reasonable accuracy overall. In contrast, the sensing-only model performs poorly

Table 2: Performance across multiple time scales. NormWear-2 achieves the best overall performance by maintaining a consistently strong trade-off across all metrics.

Model	MAE ↓	Soft-DTW ↓	FreqCos_sim ↑	FreqEucl ↓	LatentCosSim ↑	LatentEucl ↓	**Final Score ↓**
Millisecond Level (VitalDB, wearable and medical device sensing)							
Naive	0.772	814.595	0.105	1393.951	0.497	234.204	0.769
Seasonal Naive	0.909	608.661	0.393	1581.476	0.825	115.844	0.661
Sundial	0.778	655.847	0.491	1316.141	0.630	177.306	0.633
Panda	0.810	525.939	0.304	1360.377	0.647	171.515	0.652
TiReX	0.739	320.896	**0.644**	**1072.251**	0.809	116.183	0.465
Chronos-2	**0.698**	501.283	0.617	1123.497	0.808	115.609	0.500
NormWear-2 (Ours)	0.842	167.303	0.608	1134.325	0.877	94.520	**0.457**
NormWear-2 Insight only	0.883	**145.560**	0.600	1161.489	**0.894**	**86.262**	0.461
Minute Level (PMData, sport wristband daily sensing)							
Naive	0.585	460.738	0.417	429.503	0.650	219.015	0.606
Seasonal Naive	0.685	443.865	0.386	516.992	0.666	196.836	0.657
Sundial	0.609	428.150	**0.700**	416.094	0.724	175.686	0.527
Panda	0.642	459.234	0.637	430.185	0.717	175.398	0.558
TiReX	**0.477**	438.614	0.526	386.534	0.733	183.006	0.523
Chronos-2	0.481	443.463	0.493	398.692	0.716	190.676	0.541
NormWear-2 (Ours)	0.653	**293.541**	0.652	**361.957**	**0.801**	141.341	**0.466**
NormWear-2 Insight only	0.705	332.162	0.646	381.700	**0.801**	**140.976**	0.494
Minute Level (CGMacros, biofluidic sensing)							
Naive	0.761	622.353	0.520	524.629	0.609	230.554	0.709
Seasonal Naive	0.959	681.113	0.466	605.890	0.703	192.298	0.778
Sundial	0.763	479.786	0.731	458.067	0.684	193.140	0.590
Panda	0.807	465.601	0.651	441.006	0.700	181.970	0.594
TiReX	0.720	491.055	0.665	423.279	0.722	181.486	0.571
Chronos-2	**0.676**	453.967	0.683	414.460	0.723	181.411	0.548
NormWear-2 (Ours)	0.851	**239.749**	0.740	**377.904**	0.822	131.524	**0.474**
NormWear-2 Insight only	0.928	263.617	**0.743**	383.388	**0.830**	**127.924**	0.492
Quarter Hour Level (Shanghai Diabetes, biofluidic sensing)							
Naive	0.875	246.373	0.514	157.369	0.744	204.183	0.801
Seasonal Naive	0.938	143.211	0.673	130.531	0.867	126.337	0.611
Sundial	0.884	207.405	0.692	139.755	0.794	172.074	0.693
Panda	0.953	154.500	0.692	125.636	0.846	140.096	0.618
TiReX	**0.817**	180.460	0.731	124.321	0.822	155.455	0.617
Chronos-2	0.856	218.798	**0.759**	130.564	0.842	144.237	0.657
NormWear-2 (Ours)	1.008	**129.214**	0.725	**118.784**	0.880	119.986	**0.578**
NormWear-2 Insight only	1.090	144.634	0.726	120.966	**0.882**	**119.306**	0.608
Hour Level (KidneyDialysis, medical device sensing)							
Naive	0.832	136.564	0.508	102.106	0.749	211.231	0.752
Seasonal Naive	1.037	114.633	0.660	87.619	**0.847**	**148.814**	0.665
Sundial	0.846	126.489	0.729	90.928	0.791	182.420	0.662
Panda	0.867	104.077	**0.743**	80.485	0.814	168.567	0.600
TiReX	**0.808**	109.522	0.725	80.086	0.812	170.383	0.600
Chronos-2	0.835	104.425	0.748	80.286	0.832	158.743	0.589
NormWear-2 (Ours)	0.886	87.886	0.741	79.923	0.834	156.169	0.575
NormWear-2 Insight only	0.921	**87.623**	0.741	**78.217**	0.839	152.939	**0.574**

across all regimes (its error is scaled by 50% in the figure for visibility). T-SNE visualization further shows that the chaos-balanced model clearly separates systems with high and low Lyapunov exponents in latent space. These results suggest that chaos-balanced pre-training improves the model's understanding of dynamical structure and chaotic transitions, leading to better forecasting quality.

4.2 Quantitative Results of Forecasting on Physiological Signals

Table 2 summarize the main forecasting results on physiological signals under our multidimensional evaluation framework. To contextualize performance, we compare against several baselines, including running-average prediction (Naive), repetition of the dominant periodic pattern (Seasonal Naive), state-of-the-art time-series foundation models (Liu et al., 2025; Auer et al., 2026; Ansari et al., 2025), and the state-of-the-art dynamical-system foundation model (Lai et al., 2026). The digital healthcare downstream inference results shown in Table 15 primarily serve as an assessment of representation quality, providing additional context on whether models that achieve stronger generative forecasting performance also preserve the quality of their learned representations for downstream tasks. Because forecasting quality is evaluated from multiple complementary perspectives, no single method consistently achieves the best score on every individual metric. Instead, different methods exhibit distinct strengths under different criteria, underscoring the importance of multi-faceted evaluation. NormWear-2 achieves the strongest overall performance by offering the most favorable trade-off across metrics. To assess the robustness of this advantage, we conduct Conover post hoc tests on model rankings across temporal resolutions, sensing modalities, evaluation met-

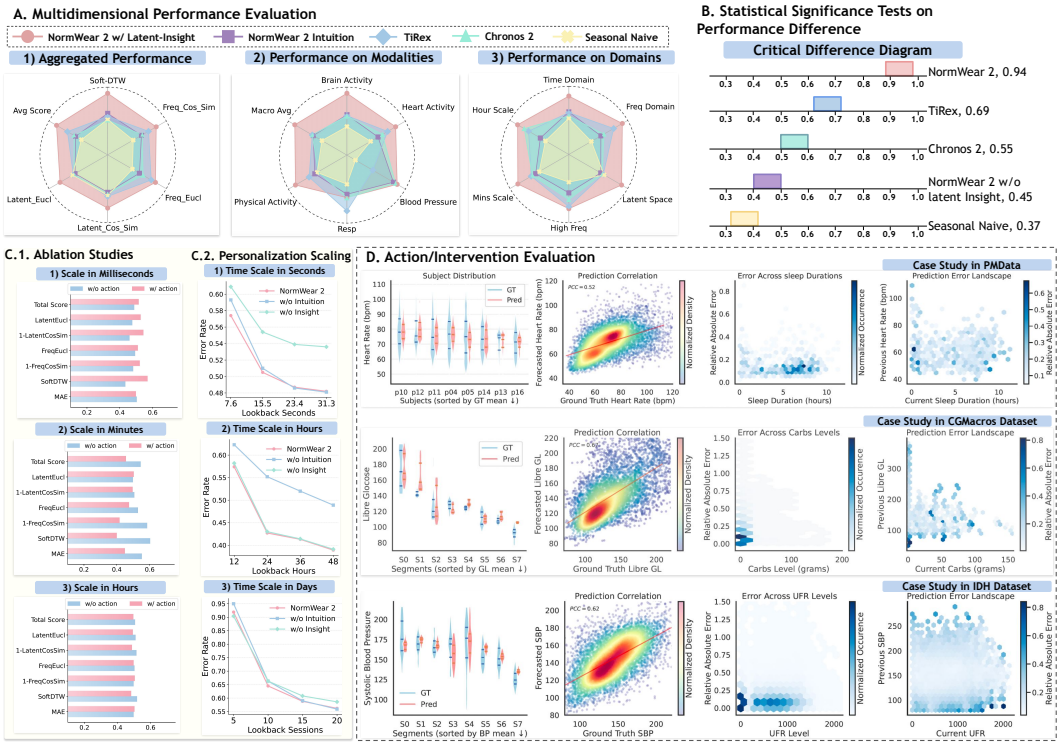

Figure 3: **Quantitative Results.** (A) Relative performance comparison of generative forecasting quality. (B) Statistical test result shows that the models performances are significantly different. (C) Overview of the ablation study results. (D) Model inference behavior under varied actions.

rics, and datasets. As shown in Figure 3A-B, NormWear-2 maintains the best overall ranking with statistical significance over prior state-of-the-art baselines.

4.3 Counterfactual Validation on Model's Action Awareness

To qualitatively assess whether the learned transition model captures action-conditioned physiological dynamics, we conduct analysis on the healthcare datasets. In the kidney dialysis scenario for example, the action variables correspond to dialysis machine control parameters prescribed to the patient and adjusted by clinicians during treatment when intolerance symptoms arise (e.g., headache, dyspnea, or chest discomfort). Among these parameters, ultrafiltration rate (UFR) is the primary intervention variable in practice. In severe cases, UFR may be reduced substantially or completely shut down to pause fluid removal. Clinically, lower UFR generally reduces the risk of adverse symptoms, but also prolongs the dialysis session, creating an inherent treatment trade-off. While optimizing dialysis control policy is beyond the scope of this work and would require a dedicated reinforcement learning framework, our objective here is to model physiological state transitions conditioned on observed clinical actions. In particular, we focus on forecasting future systolic blood pressure (SBP), one of the most critical biomarkers monitored during dialysis. Figure 3.D. provides qualitative evidence that the model learns clinically meaningful action-conditioned dynamics. First, the predicted SBP distribution remains well aligned with the ground-truth physiological range, indicating stable transition modeling. Second, predicted SBP exhibits a moderate positive linear relationship with the observed next-step SBP across representative cases, suggesting that the model preserves subject-specific physiological trends. Most importantly, when evaluated across varying UFR settings, the model maintains low relative prediction error in the majority of cases, indicating robust transition estimation under different treatment configurations.

To further inspect action sensitivity, we visualize the prediction error landscape across varying UFR levels and prior SBP conditions. Across most evaluated cases, prediction error remains consistently low and is dominated by regions below 0.1, suggesting that the learned dynamics remain stable under diverse action-state combinations. Notably, several subjects exhibit localized regions of elevated error, reflecting more complex or atypical physiological responses that are harder to capture with

Table 3: Evaluate compatibility with representative alternative backbone models. Scores are reported as (Score w/o latent insight, Score w/ latent insight). Improvements are highlighted in bold when significantly better than the non–latent-insight method. "Attn" stands for Attention.

Backbone Alternatives	MAE ↓		Soft-DTW ↓		FreqCosSim ↑		FreqEucl ↓		LatentCosSim ↑		LatentEucl ↓		**Final Score ↓**	
Millisecond Level (VitalDB, wearable and medical device sensing)														
NormWear-2 Univariate	0.888	**0.843**	301.595	**233.677**	0.475	**0.525**	1412.252	**1251.776**	0.812	0.767	116.814	134.193	0.569	**0.528**
NormWear-2 [CLS]-Attn	0.879	**0.851**	607.640	**198.523**	0.452	**0.611**	1387.582	**1138.682**	0.716	**0.860**	151.818	**101.977**	0.640	**0.468**
NormWear-2 LeWM-JEPA	0.813	0.819	298.418	**283.110**	0.525	**0.535**	1220.297	**1206.771**	0.753	0.743	139.283	142.702	0.531	**0.528**
NormWear-2 SFT	0.733	0.813	278.402	**191.238**	0.616	0.616	1085.561	1130.278	0.759	**0.858**	131.326	**102.616**	0.475	**0.459**
Minute Level (PMData, sport wristband daily sensing)														
NormWear-2 Univariate	0.625	0.633	390.088	**284.748**	0.581	**0.619**	390.470	**376.389**	0.734	**0.746**	172.824	**170.531**	0.527	**0.489**
NormWear-2 [CLS]-Attn	0.939	**0.714**	434.040	**260.89**	0.614	**0.711**	439.472	**366.758**	0.713	**0.777**	176.767	**154.105**	0.611	**0.468**
NormWear-2 LeWM-JEPA	0.614	0.661	297.461	**279.448**	0.688	**0.691**	373.915	380.756	0.757	0.743	156.661	160.783	0.471	0.480
Minute Level (CGMacros, biofluidic sensing)														
NormWear-2 Univariate	1.029	**0.967**	322.494	**245.383**	0.691	**0.710**	427.976	**400.790**	0.764	0.756	150.677	157.238	0.568	**0.527**
NormWear-2 [CLS]-Attn	1.069	**0.895**	438.332	**279.720**	0.681	**0.753**	502.969	**388.924**	0.707	**0.774**	177.405	**150.909**	0.653	**0.506**
NormWear-2 LeWM-JEPA	0.770	0.798	248.940	**237.531**	0.747	0.712	402.047	412.743	0.763	0.753	152.345	156.4	0.487	0.502
Quarter Hour Level (Shanghai Diabetes, biofluidic sensing)														
NormWear-2 Univariate	0.946	1.002	173.465	**130.461**	0.652	**0.724**	128.306	**117.504**	0.815	**0.883**	157.896	**119.879**	0.654	**0.576**
NormWear-2 [CLS]-Attn	0.861	0.954	189.533	**142.906**	0.733	0.697	133.683	**122.302**	0.802	**0.852**	166.080	**138.035**	0.654	**0.601**
NormWear-2 LeWM-JEPA	1.002	1.030	186.225	**135.697**	0.685	**0.701**	125.241	**118.791**	0.818	**0.860**	156.359	**132.92**	0.663	**0.598**
Hour Level (KidneyDialysis, medical device sensing)														
NormWear-2 Univariate	0.874	0.920	109.323	**81.342**	0.690	**0.741**	82.866	**76.013**	0.789	**0.847**	182.831	**151.855**	0.630	**0.559**
NormWear-2 [CLS]-Attn	0.864	0.916	111.051	**85.729**	0.723	**0.738**	83.154	**77.116**	0.802	**0.842**	180.434	**154.665**	0.623	**0.569**
NormWear-2 LeWM-JEPA	0.900	0.909	93.222	**84.988**	0.753	0.749	78.609	78.020	0.825	**0.833**	159.991	**155.750**	0.581	**0.569**
NormWear-2 SFT	0.789	0.889	97.428	**81.959**	0.764	**0.730**	86.400	**78.644**	0.818	**0.843**	167.1	**152.037**	0.589	**0.564**

population-level dynamics alone. These observations suggest that NormWear-2 captures meaningful action-dependent physiological trends and exhibits sensitivity to intervention changes across a broad range of treatment settings. At the same time, the observed high-error outlier cases highlight the importance of personalization: accurately modeling such subjects likely requires adapting the transition dynamics as more subject-specific historical data become available. Similar analysis results in PMData and CGMacros, representing daily-life and point-of-care scenarios, are also presented in Figure 3 panel D. More analysis examples are shown in Appendix E.2.

4.4 Compatibility with Alternative Backbones

As summarized in Table 3, our proposed latent dynamical transition modeling approach consistently improves performance when integrated with a diverse set of representative backbones, including the univariate backbone (Nie et al., 2023), the CLS-attention backbone (Luo et al., 2026), the LE-world model JEPA (Maes et al., 2026), and supervised fine-tune (SFT) backbone in scenarios with large population. Despite their architectural differences and varying inductive biases, equipping each backbone with our transition modeling mechanism leads to significant and consistent gains across multiple evaluation metrics. This demonstrates that the effectiveness of our approach is largely backbone-agnostic, highlighting its compatibility and generalizability in enhancing temporal representation learning across varied modeling paradigms.

5 Discussion and Conclusion

Limitations and Future Work. While our framework incorporates action-like mechanisms through state transition adaptation in latent space, it does not yet formulate these interactions within a general reinforcement learning paradigm with explicit reward optimization. In many real-world healthcare scenarios, decision-making involves optimizing long-term outcomes under uncertainty, where actions should be guided by well-defined objectives rather than implicit adaptation alone. Extending our latent world model with a generic RL framework, where policies operate over latent states and are trained to optimize clinically meaningful rewards, represents a key direction for future work.

Broader Impact. This work contributes to the development of general-purpose, predictive modeling frameworks for physiological signals, with potential applications in digital health, continuous monitoring, and personalized medicine. By moving beyond task-specific models toward latent representations that capture underlying dynamics, our approach may enable earlier detection of health deterioration, improved forecasting of disease progression, and more adaptive health management systems. More broadly, this work highlights the value of combining dynamical systems theory with modern representation learning as a foundation for next-generation intelligent healthcare systems.

Ethics Statement

This study contains applications in the field of healthcare. We ensured that all the data being used during pretraining and evaluations were either made publicly available by the original authors, or acquired from studies with IRB approval, and all these works were cited properly. Details of the datasets are specified in Appendix A.

LLM Disclosure

In this work, we used generative AI tools for light polishing on the use of words. The rest of the required disclosures either have not be assisted with by an LLM or are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pp. 72–78, 2019.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting, 2025. URL https://arxiv.org/abs/2510.15821.

Nieves Atienza, Rocio Gonzalez-Diaz, and Matteo Rucco. Persistent entropy for separating topological features from noise in vietoris-rips complexes. *Journal of Intelligent Information Systems*, 52(3):637–655, 2019.

Nieves Atienza, Rocío González-Díaz, and Manuel Soriano-Trigueros. On the stability of persistent entropy and new summary functions for topological data analysis. *Pattern Recognition*, 107: 107509, 2020.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. *Advances in Neural Information Processing Systems*, 38:57529–57580, 2026.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

Brian S Everitt and Anders Skrondal. *The Cambridge dictionary of statistics*, volume 4. Cambridge university press Cambridge, UK, 2010.

Navid Mohammadi Foumani, Chang Wei Tan, Geoffrey I Webb, and Mahsa Salehi. Improving position encoding of transformers for multivariate time series classification. *Data mining and knowledge discovery*, 38(1):22–48, 2024.

Benjamin Fox, Dung T Hoang, Joy Jiang, Pushkala Jayaraman, Ankit Parekh, Girish N Nadkarni, and Ankit Sakhuja. Physiojepa: Joint embedding representations of physiological signals for real time risk estimation in the intensive care unit. In *Machine Learning for Health 2025*, 2025.

Ian Fox, Joyce Lee, Rodica Pop-Busui, and Jenna Wiens. Deep reinforcement learning for closed-loop blood glucose control. In *Machine Learning for Healthcare Conference*, pp. 508–536. PMLR, 2020.

William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling. *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021.

Ricardo Gutierrez-Osuna, David Kerr, Bobak Mortazavi, and Anurag Das. Cgmacros: a scientific dataset for personalized nutrition and diet monitoring. *PhysioNet*, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Kun Hu, Plamen Ch Ivanov, Zhi Chen, Pedro Carpena, and H Eugene Stanley. Effect of trends on detrended fluctuation analysis. *Physical Review E*, 64(1):011114, 2001.

Zhiting Hu and Tianmin Shu. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*, 2023.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022.

Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*. Cambridge university press, 2003.

Jeffrey Lai, Anthony Bao, and William Gilpin. Panda: A pretrained forecast model for chaotic dynamics. In *International Conference on Learning Representations*, volume 2026, pp. 104651–104695, 2026.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019.

Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279, 2022.

Simon Lee, Cyrus Tanade, Hao Zhou, Juhyeon Lee, Megha Thukral, Md Sazzad Hissain Khan, Keum San Chun, Baiying Lu, Migyeong Gwak, Mehrab Bin Morshed, et al. Himae: Hierarchical masked autoencoders discover resolution-specific structure in wearable time series. In *International Conference on Learning Representations*, volume 2026, pp. 64859–64895, 2026.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *International Conference on Machine Learning*, 2025.

Yunfei Luo, Siwei Zhao, Subhasis Dasgupta, Tauhidur Rahman, and Rakesh Malhotra. Real-time forecasting of intradialytic hypotension using deep learning and multimodal data integration: Sapo405. *Journal of the American Society of Nephrology*, 35(10S):10–1681, 2024.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward foundation model for multivariate wearable sensing of physiological signals. *ACM Transactions on Computing for Healthcare*, 7(3):1–43, 2026.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Lalo Magni, Davide M Raimondo, Luca Bossi, Chiara Dalla Man, Giuseppe De Nicolao, Boris Kovatchev, and Claudio Cobelli. Model predictive control of type 1 diabetes: an in silico trial, 2007.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *JAMIA open*, 8(5):ooaf122, 2025.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-jepa: Learning world models through object-level latent interventions. *arXiv preprint arXiv:2602.11389*, 2026.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. *International Conference on Learning Representations*, 2024.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Chapman and Hall/CRC, 2024.

Ruifeng Tan, Weixiang Hong, Jiayue Tang, Xibin Lu, Ruijun Ma, Xiang Zheng, Jia Li, Jiaqiang Huang, and Tong-Yi Zhang. Batterylife: A comprehensive dataset and benchmark for battery life prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5789–5800, 2025.

Vajira Thambawita, Steven Alexander Hicks, Hanna Borgli, Håkon Kvale Stensland, Debesh Jha, Martin Kristoffer Svensen, Svein-Arne Pettersen, Dag Johansen, Håvard Dagenborg Johansen, Susann Dahl Pettersen, Simon Nordvang, Sigurd Pedersen, Anders Gjerdrum, Tor-Morten Grønli, Per Morten Fredriksen, Ragnhild Eg, Kjeld Hansen, Siri Fagernes, Christine Claudi, Andreas Biørn-Hansen, Duc Tien Dang Nguyen, Tomas Kupka, Hugo Lewi Hammer, Ramesh Jain, Michael Alexander Riegler, and Pål Halvorsen. Pmdata: A sports logging dataset. In *Proceedings of the 11th ACM Multimedia Systems Conference*, MMSys '20, pp. 231–236, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3339825.3394926.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *International Conference on Learning Representations*, 2025.

Rongrong Wang, Yuhu Cheng, and Xuesong Wang. Clustering-driven state embedding for reinforcement learning under visual distractions. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(12):7382–7395, 2024. doi: 10.1109/TSMC.2024.3449294.

Yifan Wang, Yijia Ma, Wen Li, and Chenyu You. Let eeg models learn eeg. In *International Conference on Machine Learning*, 2026.

Alan Wolf, Jack B Swift, Harry L Swinney, and John A Vastano. Determining lyapunov exponents from a time series. *Physica D: nonlinear phenomena*, 16(3):285–317, 1985.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *International Conference on Machine Learning*, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Qinpei Zhao, Jinhao Zhu, Xuan Shen, Chuwen Lin, Yinjia Zhang, Yuxiang Liang, Baige Cao, Jiangfeng Li, Xiang Liu, Weixiong Rao, et al. Chinese diabetes datasets for data-driven machine learning. *Scientific Data*, 10(1):35, 2023.

A DATASETS

Table 4: Lengths of input context and forecast prediction for evaluations.

Dataset	Max Observed Context	Min Observed Context	Prediction Length	# Test Samples
VitalDB (Lee et al., 2022)	2,032 (31.3 seconds)	496 (7.6 seconds)	2,064 (31.8 seconds)	62,910
PMData (Thambawita et al., 2020)	2,880 (2 days)	720 (half a day)	720 (half a day)	886
CGMacros (Gutierrez-Osuna et al., 2025)	2,880 (2 days)	720 (half a day)	720 (half a day)	173
Shanghai Diabetes (Zhao et al., 2023)	192 (2 days)	96 (1 day)	192 (2 days)	223
KidneyDialysis (Luo et al., 2024)	186 measurements (20 dialysis session)	46 measurements (5 dialysis session)	37 measurements (4 dialysis session)	22,811

Table 5: **Characteristic of the Analyzed Population in the kidney dialysis (KidneyDialysis) dataset.** SD, standard deviation; BMI, body mass index; min, minute(s).

Basic Statistics	
Total number of patients	1,452
Age (years), mean (SD)	65 (15)
BMI (Kg/m^2), mean (SD)	28.8 (8.4)
Dialysis vintage, years, mean (SD)	2.3 (1.6)
Gender, n (%)	
Male	846 (58)
Female	606 (42)
Race, n (%)	
White	789 (54)
African American	247 (17)
Hispanic	174 (12)
Number measurements per session, mean (SD)	9.3 (1.2)
Measurement gap, mean (SD), minutes	26.3 (14.7)
Measurement gap distribution (5%, 25%, 50%, 75%, 95%), min	(3, 17, 30, 31, 53)
Total number of sessions	211,397
Total number of samples	1,965,389
Validation Statistics	
Number of subjects (for_SFT_study - testing)	1,161 - 291

VitalDB is an open-access dataset designed to support machine learning research in anesthesia and perioperative monitoring. It contains high-resolution waveform and numeric biosignal data collected from 6,388 surgical cases, covering 196 intraoperative monitoring parameters, 73 perioperative clinical variables, and 34 laboratory time-series parameters. The dataset was created to address the shortage of large-scale biosignal datasets for developing predictive and analytical models in anesthesiology and patient monitoring. The detailed data statistics were presented in the original paper (Lee et al., 2022). During the evaluation, data from 20% (1,278/6,388) of total surgery cases are left out for methods assessment, and the rest of the cases are reserved for supervised fine-tune (SFT) study.

PMData is a multimodal dataset that combines daily lifelogging information with sports and physical activity records to support health and performance analysis. Data were collected from 16 participants over five months using Fitbit Versa 2 smartwatches, the PMSys smartphone application, and Google Forms. It enables research on relationships between exercise, sleep, body weight, and athletic performance, allowing both lifestyle prediction and sports-oriented analytics. The detailed data statistics were presented in the original paper (Thambawita et al., 2020).

CGMacros is a multimodal dataset developed for personalized nutrition and glucose response analysis. It includes continuous glucose monitor data, food macronutrient records, meal photographs, physical activity, demographic information, blood biomarkers, and gut microbiome profiles from 45 participants with different metabolic conditions. Participants were monitored for ten consecutive days in free-living settings while consuming meals with controlled macronutrient compositions. The detailed data statistics were presented in the original paper (Gutierrez-Osuna et al., 2025).

Shanghai Diabetes consists of two publicly available datasets created to support data-driven diabetes management research for both Type 1 and Type 2 diabetes patients. It includes data from 12 Type 1 diabetes patients and 100 Type 2 diabetes patients collected under real-life conditions in Shanghai, China. The dataset provides clinical characteristics, laboratory measurements, medications, continuous glucose monitoring records, and daily dietary information for developing glucose prediction and disease management models. The detailed data statistics were presented in the original paper (Zhao et al., 2023).

KidneyDialysis (Luo et al., 2024) is a retrospective real-world dataset, with valid IRB approval, developed for studying physiological and treatment-related patterns during maintenance hemodialysis. The dataset is acquired through a signed consent from the original authors. It includes 1,452 patients who received in-center hemodialysis across 17 Sanderling Care clinics in the United States between 2013 and 2024, comprising a total of 211,397 dialysis treatment sessions. High-resolution physiological and dialysis machine data were automatically transmitted from dialysis machines to the secure PEARL data management platform, including heart rate, body temperature, blood pressure, blood flow rate, dialysate parameters, ultrafiltration rate, cumulative fluid removal, and treatment duration. Because the original full data statistics were not reported in the original poster, we present the detailed data statistics in Table 5 based on our own analysis process. During the evaluation, data from 20% (291/1,452) of total patients are left out for methods assessment, and the rest of the cases are reserved are reserved for supervised fine-tune (SFT) study.

A.1 Pretrain Data

We leverage the datasets released by Luo et al. (2026) for 18 wearable downstream tasks, Lai et al. (2026)'s evaluation set for evaluating the generative quality on real-world chaotic system, Tan et al. (2025)'s datasets for the evaluation on battery test time series, and Wu et al. (2021)'s datasets for evaluation on civil monitoring forecasting tasks. The pre-train datasets are completely disjoint collections of multivariate time series data, with statistics presented in Table 6.

Table 6: **Pretrain Datasets.**

Datasets	Sequence Length	# Samples	# Variates
Luo et al. (2026)	390	2.3×10^5	{2,3,4,6}
Lai et al. (2026)	4096	1.0×10^5	{3,4,6}
Aggregated Benchmark	{390, 4096}	2.2×10^5	{2,3,4,6}
Balanced Benchmark	{390, 4096}	1.0×10^5	{2,3,4,6}

A.2 Evaluation Datasets for Preliminary Chaos-Aware Pre-train Experiments

The detailed statistics of the downstream tasks including the battery state of health (SOH) prediction and digital health downstream tasks were presented in prior works in the literature: Tan et al. (2025) and Luo et al. (2026) respectively. The data statistics of the genrative tasks are summarized in Table 7 as shown below.

Table 7: Statistics of generative testing datasets used in our experiments.

Dataset	Data Points per Channel	Channels	Domain
Panda Test Set (Lai et al., 2026)	3.4×10^7	[3, 6]	Chaotic System
WESAD (Luo et al., 2026)	4.2×10^6	[6]	Wearable Sensing
HUST (Tan et al., 2025)	1.0×10^6	[3]	Battery Health
CALB (Tan et al., 2025)	1.6×10^4	[3]	Battery Health
Na-ion (Tan et al., 2025)	2.0×10^4	[3]	Battery Health
Zn-coin (Tan et al., 2025)	6.4×10^5	[3]	Battery Health
ETT (Liu et al., 2025)	1.7×10^5	[7]	Civil Monitoring
Weather (Liu et al., 2025)	5.2×10^4	[21]	Civil Monitoring
illness (Liu et al., 2025)	1.0×10^3	[7]	Civil Monitoring
Exchange Rate (Liu et al., 2025)	7.2×10^3	[8]	Civil Monitoring

B Model Implementation Detail

We provide a formal description of the masking, encoding, and reconstruction pipeline used in NormWear-2. Let the input multivariate time series be $X \in \mathbb{R}^{C \times T}$ where C denotes the number of channels and T the sequence length.

B.1 Patchification

Each channel is divided into non-overlapping patches of length P:

$$X^{(c)} \to \{x_1^{(c)}, x_2^{(c)}, \ldots, x_L^{(c)}\}, \qquad x_i^{(c)} \in \mathbb{R}^P, \tag{4}$$

where $L = T/P$. Each patch is projected to a D-dimensional embedding through a Conv1D patch embedding module $E(\cdot)$:

$$z_i^{(c)} = E\big(x_i^{(c)}\big) \in \mathbb{R}^D. \tag{5}$$

where D is the latent dimension.

B.2 Channel-wise Mask Sampling

Masks are then sampled independently for each channel and each patch:

$$m_i^{(c)} \sim \text{Bernoulli}(p_{\text{mask}}), \tag{6}$$

where $m_i^{(c)}$ is the result of the indicator function with success or failure outcome sampled from the Bernoulli distribution. The masked embedding is then formed as

$$\tilde{z}_i^{(c)} = \big(1 - m_i^{(c)}\big) \cdot z_i^{(c)} \; + \; m_i^{(c)} \cdot [\text{MASK}]. \tag{7}$$

where [MASK] is a single trainable vector similar to the [CLS] special token. This produces patch-level masked embeddings for all channels. Then these embeddings are sending to the core backbone as described in method section 3.4. Denote the output from the backbone as $\hat{z}_{i,latent}^{(c)} \in \mathbb{R}^D$.

B.3 Patch-Level Deconvolution

The latent embeddings generated from the backbone are then projected back to waveform patches through a lightweight DeConv1D module:

$$\hat{s}_i^{(c)} = \text{DeConv1D}\Big(\hat{z}_{i,latent}^{(c)}\Big) \in \mathbb{R}^{P \times K} \tag{8}$$

Such a deconvolve step generates multiple candidate reconstructions $\{\hat{s}_i^{(c,k)}\}_{k=1}^K$ where K is the maximum number of candidates. These candidate reconstructions are then consolidated via a Conv1D module:

$$\hat{s}_i^{(c)} = \text{Conv1D}\Big(\hat{s}_i^{(c,1)}, \ldots, \hat{s}_i^{(c,K)}\Big) \in \mathbb{R}^{P \times 1} \tag{9}$$

and collectively, the final reconstruction is:

$$\hat{X}_{\text{reconstruct}} = \text{ConcatReshape}(\hat{s}_i^{(c)}, \;\; \forall c \in \{1, ..., C\}, i \in \{1, ..., L\}) \in \mathbb{R}^{C \times T}. \tag{10}$$

B.4 Model and Training Configuration

NormWear-2 is derived from the Masked Autoencoder (MAE) (He et al., 2022). The detailed hyper-parameter choice is describe in 8. We use a Conv1D layer with a kernel size of 16 and a stride of 16, ensuring no overlapping patches. This layer takes input with 1 channels and projects it to 768 channels, matching the hidden size of our encoders. In NormWear-2, we apply random masking independently to each variate along both the frequency and time axes, with respective masking ratios of 0.5. The masked patch are replaced by a trainable mask special token before passing to the encoder. To enhance representation learning, following Luo et al. (2026), we introduce six additional transformer blocks as fusion layers, interleaved with the original 12 encoder blocks, creating a total of 18 blocks. Each transformer block has a hidden dimension of 768 and uses LayerNorm as in

Table 8: **Pretraining Hyper-parameters.**

Hyper-parameter	Value
# cross-patches Transformer Encoder	12
# cross-channels Transformer Encoder	6
# Transformer Decoder	2
# Attention Heads	12
Encoder Latent Size	768
Decoder Latent Size	512
Feedforward Latent Size	3072
Normalization	LayerNorm
Patch size	16
Optimizer	AdamW
Loss Scalar	NativeScaler
Base Learning Rate (blr)	5e-4
Epochs	100
Batch size	128

the original MAE. The latent embeddings obtained from the encoder are projected from 768 to 512 dimensions before passing to the decoding blocks. The positional embeddings are added to guide the decoder in reconstructing the input series. The lightweight decoder consists of two transformer blocks with a hidden dimension of 512, followed by two Conv1D layers. The first Conv1D layer maps from the flattened multivariate signal embedding to an intermediate dimension, and the second Conv1D layer maps from this intermediate dimension back to the original multivariate signal space. A GELU activation function is used between these layers, with BatchNorm applied to the input. The decoder reconstructs the original input series, and the model is trained using Mean Squared Error (MSE) loss on all data points.

The models are pretrained on 8 NVIDIA RTX 3090 graphical computing unit (GPU), with 24GB of GPU memory on each card, along with 32 CPU cores, and 64GB of RAM. All the evaluation, analysis, visualization are conducted on a separate machine, with 1 NVIDIA RTX 4090 GPU, 24GB of GPU memory, 32 CPU cores, and 64GB of RAM.

B.5 Complexity Analysis of Varied Channel-Aware Encoding Mechanism

Among the three core pre-trained model comprised in this study, they leverage similar idea but different implementation to approach the multivariate time series modeling. SUNDIAL (Liu et al., 2025) propose *single-series sequence* (S3) for which they aggregate all the input series channels into a single channel along the temporal axis. When passing the time series in S3 format to the backbone model, the attention-based encoding schema is equivalent to the *All-Attention* mechanism discussed in Luo et al. (2026), which, as the authors stated, have the complexity of

$$M_{Sundial} = O(d \cdot (L \cdot C)^2) \tag{11}$$

where d is the embedding size, L as the sequence length, and C is the number of input variate or series channel. Such running complexity from the S3 approach scaled up with the production of between sequence length and the number of input channels in a polynomial manner. In comparison, NORMWEAR's final optimal channel-aware encoding schema (Luo et al., 2026), which is also leveraged in this study, have the complexity of

$$M_{NormWear} = O(d \cdot C^2) \tag{12}$$

which only scale up with the input number of variates. Lastly, PANDA (Lai et al., 2026) proposed a very similar channel-aware mechanism, which the self-attention is applied on the input variate dimension across data representation from all time series. Such design has complexity equivalent to the *Cross-Attention* as analyzed by Luo et al. (2026):

$$M_{Panda} = O(d \cdot L \cdot C^2) \tag{13}$$

Finally, we can conclude that

$$M_{Sundial} > M_{Panda} > M_{NormWear} \tag{14}$$

B.6 Action Planning Configuration

This section lists the configuration for the receding-horizon MPC experiments in Section D.7. The world model is frozen and identical to the forecasting setting. Table 9 summarizes the hyper-parameters.

Table 9: Hyper-parameter configuration for the world-model-based MPC planner and FQE evaluation.

Component	Hyper-parameter	CGMacros	ShanghaiDiabetes	KidneyDialysis
Latent MDP	K (# states)	8	4	16
	A (# actions)	4	2	8
MPC planner	H (planning horizon)	3584 (≈2.5 days)	368 (≈3.8 days)	4 (≈ 1 session)
	γ (discount factor)		0.95	
	N (sampled action seq.)		256	
	M (state transition roll-outs)		8	
FQE critic	Hidden size		256	
	Learning rate		0.001	
	Training epochs		1000	
	Batch size		256	
Protocol	random seeds		{0, 42, 100, 123, 3407}	

C Evaluating Data Balance

To evaluate the extent of data balance in terms of the metrics from chaotic theory as leveraged in this study, we mainly consider two main aspects, namely homogeneity and granularity of the distribution as presented in Figure 2 section A. Varied approaches for inspecting the balance attributes are presented below.

C.1 Unnormalized Shannon Entropy.

Shannon entropy (Shannon, 1948) is one of the metrics widely used to evaluate the amount of information within a probability distribution:

$$H(p) = -\sum_{i=1}^{n} p_i \log(p_i) \tag{15}$$

where p denote a set of probability sum up to 1. Such entropy value not only reflect homogeneity of a distribution, but also comprise granularity information, which is indicated by the fact that the more group of system that can be clustered from a dataset, the more likely the higher the value of $H(p)$.

C.2 Weighted sum of normalized Shannon Entropy and Granularity.

Since normalized Shannon entropy (with denominator of $\log(n)$) is often used under different scenario, we need to have a separate metric to explicitly evaluate the extent of granularity of a distribution. To achieve this, we leveraged a straightforward scoring formula that reflect the relative granularity:

$$G(p) = \frac{|p|}{\max_{p' \in P} |p'|} \tag{16}$$

where P is the collection of all the distributions in comparison, and $|p|$ indicate the number of bins or possible outcome of a particular distribution. The final balance score $B(p)$ is then expressed as a weighted sum of:

$$B(p) = \alpha \cdot \frac{H(p)}{\log(|p|)} + (1 - \alpha) \cdot G(p) \tag{17}$$

Since the value range of $H(p)$ and $G(p)$ is different, with relative ratio of total scores of around $4:6$, we use $\alpha = 0.6$ to balance this metric.

C.3 Weighted sum of Coefficient of Variation and Granularity.

Another commonly used metric that also reflect the extent of homogeneity of a distribution is coefficient of variation (Everitt & Skrondal, 2010), which is built on an assumption that the distribution is approximately a Gaussian distribution. From Figure 2 panel A, we observed that most of the distribution is nearly a Gaussian distribution with different mean and variance. We then leverage this new metric similar to the metric in section C.2, with $\frac{H(p)}{\log(|p|)}$ replaced by $\frac{1}{CV(p)}$, where $CV(p)$ is defined by:

$$CV(p) = \frac{\sigma_p}{\mu_p} \tag{18}$$

Similarly, we use $\alpha = 0.5$ for the same reason in section C.2 to balance the metric.

C.4 Procedure of inspect extent of balance using chaos theory based metrics

The balance inspection involve 3 main stages: (I) metrics computation, (II) unsupervised clustering, and (III) balance score computation. The first stage generally follows the procedure described in algorithm 2, which is implemented with multi-process manner as demonstrated in the codebase. The second stage follows the entirely automated pipeline described in algorithm 1, where the outcome is visualized in section E.1. Finally, the balance score is computed based on the outcome from the previous two stages, where the scoring details is presented in section C. In addition, the pseudo-code for offline balancing using the chaos metrics is presented in algorithm 3.

Algorithm 1 Clustering and Dynamical-System Typing Based on Chaos Metrics

Input: Chaos feature matrix $\mathbf{F} \in \mathbb{R}^{N \times M}$

Step 1: Determine optimal cluster number.
Compute K-means for $k = 2, 3, \ldots, K_{\max}$.
Record final inertia value for each k.
Select $k^\star \gg k_{\text{optimal}}$, where k_{optimal} is identified using the elbow rule.

Step 2: Clustering
Fit K-means with $k^\star$ and obtain centroids $\{c_1, \ldots, c_{k^\star}\}$.

Step 3: Assign semantic labels to centroids.
Compute global feature means μ across all centroids.

for $j = 1$ **to** $k^\star$ **do**
  Initialize label string L_j.
  // DFA (correlation or stationarity):
  If $c_j[\mathrm{DFA}] < 0.5$, append "Anti-corr" to L_j.
  Else if $c_j[\mathrm{DFA}] < 1.0$, append "Positive-corr" to L_j.
  Else, append "Non-station" to L_j.
  // Lyapunov exponent (degree of chaos):
  If $c_j[\lambda] < 0$, append "Stable" to L_j.
  Else if $c_j[\lambda] < \mu[\lambda]$, append "Rel Chaos" to L_j.
  Else, append "Rel Very Chaos" to L_j.
  // Persistent entropy (topological complexity):
  If $c_j[\mathrm{PE}_{H0}] < \mu[\mathrm{PE}_{H0}]$, append "Low Connect Complex" to L_j.
  Else, append "High Connect Complex" to L_j.
  If $c_j[\mathrm{PE}_{H1}] < \mu[\mathrm{PE}_{H1}]$, append "Low Loop Complex" to L_j.
  Else, append "High Loop Complex" to L_j.
  Assign L_j as the type of centroid c_j.
end for

Step 4: Merge clusters with identical type labels.
Group centroids sharing the same label and compute histogram.

Return: merged cluster types and histogram

Algorithm 2 Compute Nonlinear Dynamics Metrics

Input: Pretraining dataset $\mathcal{D}$ of multichannel time series
Output: NLD metrics per channel: DFA, Persistence Entropy, Lyapunov Exponent

for each sample $X \in \mathcal{D}$ **do**
  for each channel x in X **do**
    DFA:
    $d \leftarrow \mathrm{DFA}(x)$
    Persistence Entropy:
    $\tilde{x} \leftarrow \mathrm{TakensEmbed}(x;\ \text{delay} = 1,\ \text{dim} = 5)$
    $D \leftarrow \mathrm{VietorisRipsPersistenceDiagram}(\tilde{x})$
    $p \leftarrow \mathrm{PersistenceEntropy}(D)$
    Lyapunov Exponent:
    $\lambda \leftarrow \mathrm{LyapunovExponent}(x;\ \text{embdim} = 10,\ \tau = 1,\ \text{minsep} = 10)$
    store (d, p, λ)
  end for
end for

Algorithm 3 Iterative Chaos-Balance-Aware Sampling

Input: Dataset $\mathcal{D}$ consisting of M data sources $\{\mathcal{D}_1, \ldots, \mathcal{D}_M\}$
Shuffle samples within each data source $\mathcal{D}_m$
Initialize selected set $\mathcal{S} \leftarrow \emptyset$
Initialize remaining samples $\mathcal{R}_m \leftarrow \mathcal{D}_m$ for all m
while $|\mathcal{S}| < 0.5 \times |\mathcal{D}|$ **do**
  for $m = 1$ to M **do**
    if $\mathcal{R}_m \neq \emptyset$ **then**
      Randomly sample x from $\mathcal{R}_m$
      if $\mathrm{Score}(\mathcal{S} \cup \{x\}) > \mathrm{Score}(\mathcal{S})$ **then**
        $\mathcal{S} \leftarrow \mathcal{S} \cup \{x\}$
      end if
      Remove x from $\mathcal{R}_m$
    end if
  end for
end while
Return $\mathcal{S}$

D SUPPLEMENTARY RESULTS

D.1 CAUSAL VERSUS BIDIRECTIONAL ATTENTION IN THE BACKBONE

This experiment investigates whether the masking scheme of the backbone encoder affects forecasting quality. A bidirectional encoder is trained with access to context from both directions but has access only to past context during autoregressive generation, thereby introducing a mismatch between training and inference. In contrast, causal-masked attention uses past-only context during both training and inference, thus avoiding this mismatch. We compare the two variants with all other components held fixed and without the latent insight module, such that the comparison isolates the effect of the backbone. Both variants are evaluated on the five clinical datasets using the metrics defined in Section 3; the final score (Avg) is computed according to Equation 3.

The results are reported in Table 10. The causal variant achieves a higher performance score on four of the five datasets. On KidneyDialysis, the bidirectional variant performs marginally better, although the causal variant remains competitive. We attribute this observation to the very long inter-measurement horizon of the kidney dataset (Appendix A, Table 5), which may reduce the impact of local context ordering. We therefore adopt causal masking as the default backbone in the main experiments.

Table 10: Causal versus bidirectional masking in the backbone (without latent insight). Lower Avg is better. The best Avg per dataset is in bold.

Dataset	Variant	MAE ↓	Soft-DTW ↓	FreqCos ↑	FreqEucl ↓	LatCos ↑	LatEucl ↓	Avg ↓
VitalDB	Bi-dir	0.879	607.640	0.452	1387.582	0.716	151.818	0.640
	Causal	0.898	208.931	0.478	1337.775	0.812	116.790	**0.543**
PMData	Bi-dir	0.939	434.040	0.614	439.472	0.713	176.767	0.611
	Causal	0.541	379.293	0.592	379.295	0.712	192.806	**0.511**
CGMacros	Bi-dir	1.069	438.332	0.681	502.969	0.707	177.405	0.653
	Causal	0.722	367.525	0.713	385.192	0.721	178.416	**0.517**
ShanghaiDiabetes	Bi-dir	0.861	189.533	0.733	133.683	0.802	166.080	0.654
	Causal	0.880	168.933	0.677	131.727	0.819	155.087	**0.640**
KidneyDialysis	Bi-dir	0.864	111.051	0.723	83.154	0.802	180.434	**0.623**
	Causal	0.846	111.886	0.717	88.576	0.788	184.920	0.640

D.2 CHAOS-AWARE BALANCING VERSUS SOURCE HETEROGENEITY

This experiment isolates the effect of chaos-aware balancing from that of simply adding heterogeneous pretraining sources. All three settings draw from the same pool of synthetic chaotic-system and high-frequency wearable data under a fixed sample budget. *Random Sample* samples uniformly from this pool. *Concatenation* combines all sources without any balancing. *Balancing* selects a subset according to the chaos-aware metrics described in Section 3. Performance is aggregated across the general time-series benchmarks (wearable, chaotic systems, civil monitoring, and battery) listed in Table 7, under both the forecast and simulate settings. *Short* and *Long* denote predicting the subsequent 50% and 100% of a test sequence, respectively, given its context portion. We report MAE as mean $\pm$ standard deviation. Table 11 shows that uniform random sampling achieves the strongest performance on the forecasting tasks, whereas balancing performs best on the simulation tasks and across all aggregate metrics (short, long, and micro average). Concatenation without balancing consistently underperforms balancing, suggesting that the observed gains arise from the selection criterion rather than from simply incorporating heterogeneous data.

D.3 SUPPLEMENTARY PERFORMANCE METRIC: FRÉCHET INCEPTION DISTANCE (FID)

This experiment examines whether the method ranking observed in the main evaluation remains consistent under a distribution-level metric. Following Wang et al. (2026), we compute the Fréchet

Table 11: Fixed-budget pretraining-data ablation on the same source pool. MAE (mean ± std.), lower is better. Bold marks the best and italics the second best.

Task	Random Sample	Concatenation	Balancing (Ours)
Short Forecast	**0.558 ± 0.010**	0.666 ± 0.033	*0.597 ± 0.032*
Long Forecast	**0.696 ± 0.019**	0.762 ± 0.032	*0.729 ± 0.028*
Short Simulate	0.596 ± 0.011	*0.484 ± 0.056*	**0.449 ± 0.043**
Long Simulate	0.702 ± 0.018	*0.555 ± 0.053*	**0.537 ± 0.042**
Short Average	0.577 ± 0.010	*0.575 ± 0.051*	**0.523 ± 0.041**
Long Average	0.699 ± 0.017	*0.658 ± 0.052*	**0.633 ± 0.043**
Micro Average	0.638 ± 0.017	*0.596 ± 0.052*	**0.557 ± 0.044**

Inception Distance (FID) between generated and ground-truth trajectories in two domains: the frequency domain and the latent domain, following the same construction described in Section 3. We compare the naive baseline, Sundial, Chronos-2, our full model, and our model with only the intuition module, across all five health datasets. The results are reported in Table 12. Our full model achieves the lowest FID on every dataset in both domains. Moreover, the relative ordering of all methods is consistent with the ranking obtained from the main metrics presented in Section 4.

Table 12: FID in the frequency domain (top) and latent domain (bottom); lower is better. The best result in each column is shown in bold.

Model	VitalDB	PMData	CGMacros	ShanghaiDiabetes	KidneyDialysis
Frequency domain					
Naive	1526.692	916.202	666.187	204.829	191.614
Sundial	1318.657	769.534	577.141	130.899	155.538
Chronos-2	818.432	802.333	552.554	71.089	94.346
Ours w/o insight	885.196	764.413	507.738	108.494	141.518
Ours	**484.395**	**366.974**	**249.303**	**42.763**	**89.610**
Latent domain					
Naive	3999.551	2733.668	3391.489	1786.296	2374.094
Sundial	1870.372	1261.724	1739.191	903.898	1193.290
Chronos-2	675.766	1466.788	1392.597	558.887	643.167
Ours w/o insight	593.285	1312.836	1277.630	829.929	1111.449
Ours	**448.481**	**447.375**	**396.931**	**317.529**	**508.615**

D.4 Influence on Final Score when Alternating Latent Transition Path

This experiment evaluates whether the final score is sensitive to the quality of the learned latent transition mechanism, rather than merely to reconstruction fidelity. The score combines Soft-DTW (temporal and morphological consistency), frequency-domain metrics (spectral fidelity), and latent-space metrics (representation-level agreement), thereby capturing quality across the time, frequency, and latent domains. At inference, we perturb the learned transition paths in two ways: (a) *always_low*, which forces every transition to the lowest-probability latent state and provides a strong negative control; and (b) *reverse_prob*, which inverts the transition probabilities while still assigning non-trivial probability mass to plausible states. The unperturbed model (Ours) is included as a reference. Table 13 reports the results. The *always_low* perturbation consistently degrades performance across all five datasets, as it forces trajectories into atypical, low-probability regions and thereby disrupts the learned latent transition structure. This provides direct evidence that forecasting quality depends on the validity of the learned transition space rather than on reconstruction fidelity alone.

In contrast, *reverse_prob* generally remains closer to the unperturbed model, with regular transition sampling outperforming it in most cases. In the few cases where *reverse_prob* performs better, the difference can be attributed in part to the sparsity of the observed states. For example, KidneyDialysis contains relatively long inter-measurement intervals, with some measurements such as blood pressure recorded only every half hour. When plausible next states remain available within the valid

transition region, reversing their probabilities can still preserve substantial mass on states close to the correct transition. Because near-term physiological dynamics are typically smooth, these plausible states may yield similar latent representations and forecasts despite differences in their relative ranking. This helps explain why *reverse_prob* remains competitive in these cases.

Taken together, the results demonstrate two key points. First, forecasting quality is strongly dependent on the validity of the latent transition space, as substantially disrupting this space leads to consistent performance degradation. Second, the learned transitions of our model achieve the best aggregate score on all five datasets, with the ordering *Ours* > *reverse_prob* > *always_low* consistently reflecting the increasing severity of the perturbation. This monotonic pattern further supports the conclusion that downstream forecasting quality is closely tied to the quality of the learned latent transition paths.

Table 13: Effect of perturbing the latent transition paths at inference. Lower Avg is better; the best Avg for each dataset is shown in bold.

Dataset	Method	MAE ↓	Soft-DTW ↓	FreqCos ↑	FreqEucl ↓	LatCos ↑	LatEucl ↓	Avg ↓
VitalDB	always_low	0.886	601.800	0.250	1614.598	0.663	180.620	0.726
	reverse_prob	0.895	177.222	0.538	1225.638	0.871	97.293	0.495
	Ours	0.842	167.303	0.608	1134.325	0.877	94.520	**0.457**
PMData	always_low	0.747	419.406	0.538	454.779	0.688	181.302	0.600
	reverse_prob	0.832	324.520	0.688	385.146	0.789	139.940	0.510
	Ours	0.653	293.541	0.652	361.957	0.801	141.341	**0.466**
CGMacros	alway_low	0.891	397.202	0.549	488.000	0.740	164.408	0.621
	reverse_prob	0.930	203.814	0.682	405.121	0.838	123.666	0.497
	Ours	0.851	239.749	0.740	377.904	0.822	131.524	**0.474**
Shanghai Diabetes	always_low	1.114	278.804	0.499	169.963	0.722	205.766	0.897
	reverse_prob	1.048	142.373	0.695	124.036	0.862	130.054	0.616
	Ours	1.008	129.214	0.725	118.784	0.880	119.986	**0.578**
Kidney Dialysis	always_low	0.993	109.409	0.602	92.770	0.771	187.953	0.695
	reverse_prob	1.015	89.690	0.736	77.590	0.846	147.362	0.590
	Ours	0.886	87.886	0.741	79.923	0.834	156.169	**0.575**

D.5 Personalization with Increasing Context

Table 14: Final score (lower is better) as a function of context window length across three datasets with different native time scales. The best result in each column is shown in bold.

	VitalDB (seconds)				**PMData** (hours)				**KidneyDialysis** (sessions)			
Method	7.6	15.5	23.4	31.3	12	24	36	48	5	10	15	20
Mean	0.805	0.782	0.774	0.769	0.858	0.698	0.679	0.606	1.114	0.853	0.784	0.752
Seasonal	0.831	0.803	0.793	0.661	0.954	0.777	0.763	0.657	1.187	0.918	0.820	0.665
Ours	**0.574**	**0.505**	**0.487**	**0.482**	**0.574**	**0.427**	**0.413**	**0.389**	**0.919**	**0.645**	**0.588**	**0.562**

This experiment accompanies Fig. 3.C. and evaluates whether the model supports a simple continuous-improvement personalization pipeline, in which the available per-subject history grows over time. For each of three datasets with distinct native time scales (seconds for VitalDB, hours for PMData, and dialysis sessions for KidneyDialysis), we vary the context window and report the final score defined in Equation 3, which is also plotted on the vertical axis of Fig. 3.C. We include two history-based baselines: mean prediction, which forecasts each channel using its mean value over the context, and seasonal naive, which repeats the most recent observed period.

The results are reported in Table 14. Our model improves monotonically as the context window increases across all three datasets and consistently outperforms both baselines at every window size. The baselines likewise benefit from additional context, with seasonal naive surpassing mean prediction at the longest window on each dataset. Nevertheless, our model maintains a substantial

performance advantage throughout. This trend suggests that NormWear-2 can effectively leverage accumulating user-specific history to achieve progressively better personalized forecasting, highlighting its potential for continual improvement in real-time monitoring deployments.

D.6 Inspecting Downstream Representation Quality

Table 15: Inspect the quality of learned representations on digital health applications under linear probing evaluation. The signal name, in column "Modality-Specific", following each performance score denotes the model specialized for that modality: PPG (Pillai et al., 2024), ECG-FM (McKeen et al., 2025), and EEG (Wang et al., 2025).

Downstream Tasks	Panda	Sundial	Chronos 2	TiReX	Modality-Specific	NormWear (Luo et al., 2026)	Normwear 2 (Ours)
WESAD	73.187	70.529	74.414	73.802	56.656(PPG))	**76.060**	72.524
UCI-HAR	97.896	96.522	98.476	97.887	-	**98.954**	98.141
DriverFatigue	68.116	70.034	73.551	69.447	80.430(EEG)	**74.292**	73.178
Activity Recognition Avg.	79.733	79.028	82.147	80.379	-	**83.102**	81.281
Epilepsy (eye open state)	89.958	**95.797**	94.968	94.351	90.436(EEG)	92.743	93.676
Epilepsy (eye relaxation)	94.085	97.390	**97.723**	97.420	95.552(EEG)	94.828	96.639
Epilepsy (health area)	89.047	**91.812**	91.487	90.634	88.065(EEG)	88.541	90.079
Epilepsy (tumor area)	86.415	**91.103**	90.104	90.003	87.258(EEG)	87.197	88.257
Epilepsy (seizure)	98.636	**99.723**	99.541	99.274	94.616(EEG)	97.053	99.339
GAMEEMO	55.263	**56.814**	56.795	55.194	55.420(EEG)	54.937	54.946
EEG Main Tasks Avg.	85.567	**88.773**	88.436	87.813	85.225	85.883	87.156
ECG-Abnormal	99.542	99.730	99.426	**99.737**	89.898(ECG)	99.140	98.605
PPG-BP (HTN)	56.082	55.47	55.963	61.455	61.839(PPG)	**62.341**	62.268
PPG-BP (DM)	54.992	58.033	59.234	59.647	55.668(PPG)	55.893	**62.087**
PPG-BP (CVA)	51.875	59.514	45.208	43.125	**73.125(PPG)**	70.625	70.347
PPG-BP (CVD)	63.121	59.275	**69.223**	54.685	49.066(PPG)	51.773	62.021
PhysioNet EMG	99.948	99.999	99.999	99.546	-	99.216	99.999
Risk Evaluation Avg.	70.927	72.004	71.509	69.699	-	73.165	**75.888**
Noninvasive-BP	92.907	90.857	91.995	92.346	90.596(PPG)	92.420	**93.100**
PPG-Hgb	94.451	94.419	92.844	94.862	**94.912(PPG)**	94.632	93.779
Fetal-fPCG	98.884	99.082	**99.105**	98.937	-	99.072	98.990
Vital Signs Avg.	**95.414**	94.786	94.648	95.382	-	95.375	95.290
Micro Avg.	81.356	82.561	82.781	81.797	-	82.762	**83.776**
Macro Avg.	82.910	83.648	84.185	83.318	-	84.381	**84.904**

D.7 Action Planning with the Latent World Model

To test whether the learned world model could benefit action planning in addition to plain forecasting, we construct a lightweight straightforward receding-horizon model predictive (MPC) (Chua et al., 2018) based planning process on top of the world model and evaluate the resulting policy comparing against the logged action trajectory (denoted as "behavior") on three health datasets with well defined intervention objective and reward functions: CGMacros and ShanghaiDiabetes (glucose control) and KidneyDialysis (intradialytic hypotension (IDH)). 80% of the subjects are used for action discretization and one-scan in-context learning as described as the insight inference pathway in Section 3, and the performance are reported on the rest 20% subjects by conducting MPC using the world model for state-action forecasting. Fitted Q-Evaluation (FQE) (Le et al., 2019) is leveraged as the main evaluation metric which are fitted on the entire dataset including test data.

Decision process. Each trajectory is z-normalized with training statistics and encoded by the frozen encoder. Every patch of raw steps yields one state token with patch size following the world model's configuration (channel-flattened embedding), so one decision corresponds to one patch in the latent space, and the raw value of the action is decoded through the decoder of the world model. The raw action channels is discretized into fixed action space by k-means on training data. Rewards are negative clinical risk scores computed on the raw target channel and summed over the patch. For management of glucose g (mg/dL), we use the Magni risk (Magni et al., 2007; Fox et al., 2020) as the standard reward function:

$$r_t^{\text{glu}} = -10\left(3.5505\,(\ln g_{t+1})^{0.8353} - 3.7932\right)^2, \tag{19}$$

And for IDH risk (Luo et al., 2024) following clinical suggestions, with systolic blood pressure denoted as b (mmHg), we have reward function:

$$r_t^{\text{idh}} = -\big(0.25\,\mathbb{I}[\Delta_t{>}20] + 0.4\,\mathbb{I}[\Delta_t{>}30] + \mathbb{I}[b_{t+1}{<}90] + 0.5\,\mathbb{I}[b_{t+1}{<}100]\big), \quad \Delta_t = b_t - b_{t+1}. \tag{20}$$

World model and planner. The world model provides a discrete latent MDP fit on encoded training trajectories: K k-means centroids $\{c_k\}$ act as discrete states, and an empirical Markov matrix $T \in \mathbb{R}^{K\times K}$ counts transitions within trajectories. Action conditioning re-weights each row toward states whose action-channel slice is close to the encoded action token e_a,

$$T_a(k,k') \;\propto\; T(k,k')\,\big(1-\tilde{d}(k',a)\big), \qquad \tilde{d}(k',a) = \frac{d(k',a) - \min_j d(j,a)}{\max_j d(j,a) - \min_j d(j,a)}, \tag{21}$$

with $d(k',a) = \|c_{k'}^{\text{act}} - e_a\|_2$ where e_a is the corresponding latent embedding of the action channel of the corresponding patch. Each centroid is decoded once through the frozen decoder to its raw target-channel window $w_k \in \mathbb{R}^L$, so any predicted latent path can be mapped to a predicted raw signal and scored with Eq. equation 19 or equation 20 on the fly. Given the current state quantized to k_0, the planner samples N random action sequences $\mathbf{a}^{(n)} \in \{1,\dots,A\}^H$, we simulates each state transitions M times through T_a, and selects

$$a^* \;=\; \arg\max_n \; \frac{1}{M}\sum_{m=1}^{M}\sum_{t=0}^{H-1} \gamma^t\, R\big(w_{k_t^{(n,m)}}\big), \qquad k_{t+1}^{(n,m)} \sim T_{a_t^{(n)}}\big(k_t^{(n,m)},\cdot\big), \tag{22}$$

where $R(\cdot)$ sums the per-step reward over the decoded window, and γ is the decay coefficient. Only a^* is executed at current time step and the controller re-plans at the next observed state. The horizon H is the median trajectory length in decisions γ.

Evaluation. Since the datasets are static, we use fitted Q-evaluation (FQE) (Le et al., 2019), which is a model free evaluation approach for static state-action-reward dataset. A lightweight two-layer MLP $Q(s,a)$ is regressed on the Monte-Carlo discounted reward-to-go of the logged trajectories. Because no Markov assumption on the latent state in this case, instead of using Bellman equation, the actual long-term reward $G_t = \sum_{t=0}^{H-1}\gamma^t R_t$ is leveraged, and a policy π is scored by $\frac{1}{|\mathcal{S}_0|}\sum_{s\in\mathcal{S}_0}\mathbb{E}_{a\sim\pi(\cdot|s)}Q(s,a)$ over the initial states $\mathcal{S}_0$ of held-out subjects. We compare MPC against five references evaluated with the same Q: the logged *behavior* action, a *uniform*-random action, the constant *mean* and *mode* training action, and a per-state s_0-*mode* (most frequent training action in each discrete latent state). Results are averaged over multiple random seeds.

Table 16: FQE-estimated return, reported as mean ± ±std deviation over multiple random seeds, for six action strategies across three static health datasets. Higher values are better; the best result for each dataset is shown in **bold**. MPC uses actions generated by our proposed world model.

Dataset	MPC (Ours)	Behavior	Uniform	Mean	Mode	s_0-Mode
CGMacros	**−2958.354** ±136.995	−3125.615 ±27.657	−3007.378 ±113.148	−3125.615 ±27.657	−3125.615 ±27.657	−3125.615 ±27.657
ShanghaiDiabetes	**−1673.606** ±62.920	−1721.517 ±4.882	−1954.088 ±37.889	−1785.368 ±13.338	−1785.368 ±13.338	−1723.244 ±38.158
KidneyDialysis	**−1.820** ±0.233	−1.998 ±0.159	−1.885 ±0.180	−1.869 ±0.197	−2.092 ±0.203	−2.047 ±0.215

Results. Table 16 shows that MPC on top of the world model obtains the highest FQE return on all three datasets, improving over the logged behavior policy by 5.4% (CGMacros), 2.8% (ShanghaiDiabetes), and 8.9% (KidneyDialysis), and over every non-learned reference. Since the planner uses nothing beyond the world model's transition and decoder plus a clinical risk function, this indicates that the learned latent dynamics indicate potential actionable information about how interventions could shape future physiology. Detailed hyper-parameter setting is presented in Appendix B.6. Given that this work primarily focuses on world modeling, a more in-depth investigation of action planning, an agentic modeling problem by definition, is left for future work.

D.8 SUPPLEMENTARY DATA-BALANCE-AWARE STUDY EVALUATION PERFORMANCE

Table 17 contains more fine-grained report of the performance of the methods in comparison. Table 20 present the analysis result of the horizontal drift of models' performance.

Table 17: **Preliminary Evaluation Performance.** All the generative evaluation tasks are zero-shot forecasting on multivariate time series. The averaged results across test scenarios within each domain and each task are reported. For all the generative tasks, a lower MSE or MAE indicates a better prediction (Liu et al., 2025). For Battery SoH downstream tasks, mean absolute percentage error and R^2 are reported as the main metrics (Tan et al., 2025). For the wearable downstream tasks, AUC-ROC is reported as the main metric (Luo et al., 2026). 1st Count represents the number of wins achieved by a model across all domains and test scenarios.

(a) **Generative Task Results**

Evaluation Scenario	**SUNDIAL**$_{large}$		**PANDA**		**NORMWEAR-2 (Ours)**		**NORMWEAR-2 Chaotic only**		**NORMWEAR-2 Sensor only**	
Metric	MAE ↓	MSE ↓	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Chaotic Short Forecast	0.881	1.222	**0.486**	**0.587**	0.678	0.893	0.652	0.838	0.904	1.342
Chaotic Long Forecast	0.870	1.175	**0.610**	**0.785**	0.770	1.040	0.750	0.990	0.920	1.380
Chaotic Short Simulation	0.874	1.198	**0.397**	**0.464**	0.629	0.787	0.649	0.839	0.900	1.344
Chaotic Long Simulation	0.860	1.155	**0.610**	**0.785**	0.770	1.040	0.730	0.925	0.935	1.435
Chaotic Generative Avg.	0.889	1.241	**0.510**	**0.638**	0.645	0.833	0.631	0.796	0.896	1.319
Wearable Short Forecast	0.930	2.257	0.852	2.236	**0.723**	**1.555**	0.743	1.631	0.765	1.670
Wearable Long Forecast	0.950	2.170	1.010	2.750	**0.880**	**1.950**	0.880	1.960	0.950	2.330
Wearable Short Simulation	0.924	2.109	0.855	1.885	**0.790**	**1.789**	0.793	1.821	0.840	1.932
Wearable Long Simulation	0.970	2.260	0.950	2.180	**0.900**	**2.070**	0.900	2.120	0.960	2.300
Wearable Generative Avg.	0.953	2.255	0.840	1.933	**0.740**	**1.538**	0.810	1.813	0.835	1.883
Civil Short Forecast	0.539	0.576	**0.413**	**0.405**	0.503	0.561	0.492	0.535	0.533	0.590
Civil Long Forecast	0.654	0.953	**0.557**	**0.778**	0.636	0.938	0.633	0.925	0.678	1.039
Civil Short Simulation	0.569	0.614	0.509	**0.516**	0.518	0.569	**0.508**	0.549	0.582	0.670
Civil Long Simulation	0.675	0.961	**0.619**	**0.816**	0.640	0.925	0.632	0.897	0.697	1.061
Civil Generative Avg.	0.609	0.776	**0.525**	**0.628**	0.574	0.748	0.566	0.727	0.622	0.840
Battery Short Forecast	0.926	1.311	0.616	0.600	**0.512**	**0.477**	0.633	0.695	0.951	1.306
Battery Long Forecast	0.947	1.292	0.702	0.749	**0.671**	**0.749**	0.777	0.954	1.005	1.428
Battery Short Simulation	0.791	0.967	1.053	1.628	**0.610**	**0.642**	0.707	0.770	0.977	1.377
Battery Long Simulation	0.833	0.997	0.956	1.333	**0.715**	**0.759**	0.796	0.914	1.007	1.441
Battery Generative Avg.	0.874	1.141	0.832	1.077	**0.627**	**0.657**	0.728	0.833	0.985	1.388
Short Forecast All.	0.819	1.341	**0.592**	0.957	0.604	**0.871**	0.630	0.925	0.788	1.227
Long Forecast All.	0.855	1.397	**0.720**	1.265	0.740	**1.174**	0.762	1.216	0.875	1.470
Short Simulate All.	0.789	1.222	0.704	1.123	**0.637**	**0.947**	0.664	0.995	0.825	1.331
Long Simulate All.	0.834	1.343	0.784	1.278	**0.738**	**1.149**	0.761	1.192	0.884	1.499
Generative Micro Avg.	0.780	1.085	0.668	0.915	**0.638**	**0.841**	0.672	0.900	0.828	1.234

(b) **Downstream Task Results**

Metric	MAPE ↓	R^2 ↑	MAPE	R^2	MAPE	R^2	MAPE	R^2	MAPE	R^2
Battery Li-ion Downstream	0.154	0.040	0.198	-0.726	**0.145**	**0.229**	0.202	-0.342	0.158	0.084
Battery Na-ion Downstream	0.050	0.954	0.075	0.879	**0.047**	**0.963**	0.070	0.888	0.051	0.947
Battery Zn-coin Downstream	1.550	0.551	1.321	0.255	**0.685**	**0.563**	0.964	0.447	0.695	0.371
Battery Downstream Micro Avg.	0.628	0.500	0.717	0.073	0.366	**0.531**	0.396	0.290	**0.310**	0.328
Metric	AUC ROC ↑		AUC ROC		AUC ROC		AUC ROC		AUC ROC	
Wearable State Recognition	0.790		0.797		**0.813**		0.799		0.790	
Wearable EEG Tasks	**0.888**		0.856		0.872		0.862		0.864	
Wearable Vital Sign (1-MAPE)	0.948		**0.954**		0.953		0.949		0.953	
Wearable Disease Risk	0.720		0.709		**0.759**		0.741		0.694	
Wearable Downstream Avg.	0.826		0.814		**0.838**		0.826		0.810	
1st Count	12		23		**30**		5		4	

Table 18: Generative results of on scaling in pre-train subsets with varied balance scores.

Data Size	10^3						10^4						10^5					
Evaluation Scenario	**Balance**		**Chaotic**		**Sensor**		**Balance**		**Chaotic**		**Sensor**		**Balance**		**Chaotic**		**Sensor**	
Metric	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Chaotic Generative Avg.	0.830	1.064	0.803	1.031	1.051	1.791	0.698	0.892	0.746	0.962	1.079	1.891	0.696	0.900	0.695	0.898	0.912	1.369
Wearable Generative Avg.	0.835	1.829	0.841	1.822	0.960	2.218	0.826	1.815	0.827	1.809	0.935	2.195	0.821	1.836	0.828	1.870	0.854	1.930
Civil Generative Avg.	0.608	0.786	0.595	0.762	0.731	1.125	0.565	0.717	0.575	0.738	0.715	1.105	0.574	0.748	0.566	0.727	0.622	0.840
Battery Generative Avg.	0.911	1.075	0.875	1.007	0.980	1.337	0.846	0.975	0.833	0.960	1.028	1.462	0.627	0.657	0.728	0.833	0.985	1.388
Short Forecast All.	0.752	1.067	0.740	1.050	0.910	1.557	0.674	0.933	0.690	0.969	0.898	1.505	0.604	0.871	0.630	0.925	0.788	1.227
Long Forecast All.	0.836	1.324	0.817	1.262	0.956	1.705	0.771	1.191	0.787	1.216	0.959	1.717	0.740	1.174	0.762	1.216	0.875	1.470
Short Simulate All.	0.770	1.105	0.751	1.069	0.912	1.536	0.705	1.018	0.710	1.016	0.924	1.596	0.637	0.947	0.664	0.995	0.825	1.331
Long Simulate All.	0.827	1.259	0.806	1.241	0.943	1.673	0.785	1.258	0.793	1.268	0.976	1.834	0.738	1.149	0.761	1.192	0.884	1.499
Generative Micro Avg.	0.779	1.036	0.757	0.996	0.900	1.422	0.715	0.943	0.723	0.957	0.915	1.477	0.638	0.841	0.672	0.900	0.828	1.234

Table 19: Performance comparison across models pretrained using different backbone architectures.

Models	**Chronos**$_{base}$ (Ansari et al., 2024)	**Uni-variate** (Nie et al., 2023)	**[CLS] Attn.** (Luo et al., 2026)	**Channel Attn.** (Lai et al., 2026)
Zero-shot Generative Tasks	MAE ↓	MAE ↓	MAE ↓	MAE ↓
Short-term forecast	**0.433±0.018**	0.564±0.019	0.558±0.010	0.653±0.032
Long-term forecast	**0.638±0.083**	0.694±0.022	0.696±0.019	0.747±0.030
Short-term simulate	0.451±0.018	0.569±0.012	0.596±0.011	**0.405±0.053**
Long-term simulate	0.696±0.135	0.706±0.019	0.702±0.018	**0.477±0.062**
All short-term generative	**0.442±0.017**	0.566±0.015	0.577±0.010	0.529±0.057
All long-term generative	0.667±0.105	0.700±0.020	0.699±0.017	**0.612±0.063**
All generative series	**0.554±0.072**	0.633±0.021	0.638±0.017	0.571±0.060
Battery SOH Downstream	R^2 ↑	R^2 ↑	R^2 ↑	R^2 ↑
Li-ion SOH	0.036±0.026	-0.190±0.059	**0.229±0.021**	-0.487±0.075
Na-ion SOH	**0.966±0.000**	0.525±0.255	0.963±0.0002	-1.628±3.145
Zn-coin SOH	0.433±0.017	**0.630±0.016**	0.563±0.012	0.541±0.054
Battery SOH Avg	0.507±0.173	0.297±0.189	**0.531±0.238**	-0.955±4.986
Wearable Downstream Tasks	AUC-ROC ↑	AUC-ROC↑	AUC-ROC ↑	AUC-ROC ↑
State Recognition	0.799±0.11	0.804±0.025	0.813±0.021	**0.814±0.023**
EEG Tasks	0.807±0.021	**0.875±0.027**	0.872±0.027	0.872±0.026
Vital Sign (1-MAPE)	**0.953±0.001**	0.951±0.001	**0.953±0.001**	0.949±0.001
Disease Risk	0.621±0.034	0.743±0.037	**0.759±0.034**	0.724±0.045
Wearable Avg.	0.768±0.032	0.832±0.028	**0.838±0.025**	0.826±0.031
Avg. Rank	2.250	2.688	**2.188**	2.750

Table 20: Relative Drop in MAE (%) and Average Drift across different methods between long term and short term generative series. The drift speeds for all methods are on the order of 10^{-4}, indicating that error accumulation over these horizons is minimal.

Metric	**Sundial**	**Panda**	**NormWear-2**	**NormWear-2 Chaotic only**	**NormWear-2 Sensor only**
Chaotic Drop	+0.014%	-38.165%	-13.208%	-13.610%	-2.162%
Wearable Drop	-3.560%	-14.821%	-16.920%	-15.690%	-12.710%
Civil Drop	-19.955%	-27.611%	-24.994%	-26.393%	-23.223%
Battery Drop	-3.610%	+0.659%	-23.546%	-17.491%	-4.331%
Avg Drop	-6.439%	-19.985%	-19.667%	-18.296%	-10.607%
Chaotic Drift	4.7×10^{-5}	6.6×10^{-4}	3.4×10^{-4}	3.5×10^{-4}	7.6×10^{-5}
Wearable Drift	1.3×10^{-4}	4.9×10^{-4}	5.0×10^{-4}	4.7×10^{-4}	4.0×10^{-4}
Civil Drift	4.3×10^{-4}	5.0×10^{-4}	5.0×10^{-4}	5.2×10^{-4}	5.1×10^{-4}
Battery Drift	1.2×10^{-4}	2.1×10^{-5}	5.2×10^{-4}	4.6×10^{-4}	1.6×10^{-4}
Avg Drift	1.6×10^{-4}	4.1×10^{-4}	4.6×10^{-4}	4.5×10^{-4}	2.9×10^{-4}

E QUALITATIVE VISUALIZATION

E.1 VISUALIZING CHAOS METRICS BASED CLUSTERS OF DYNAMICAL SYSTEMS

The details of the description of varied time series system in term of the chaotic theory based metrics are presented in Figure 4.

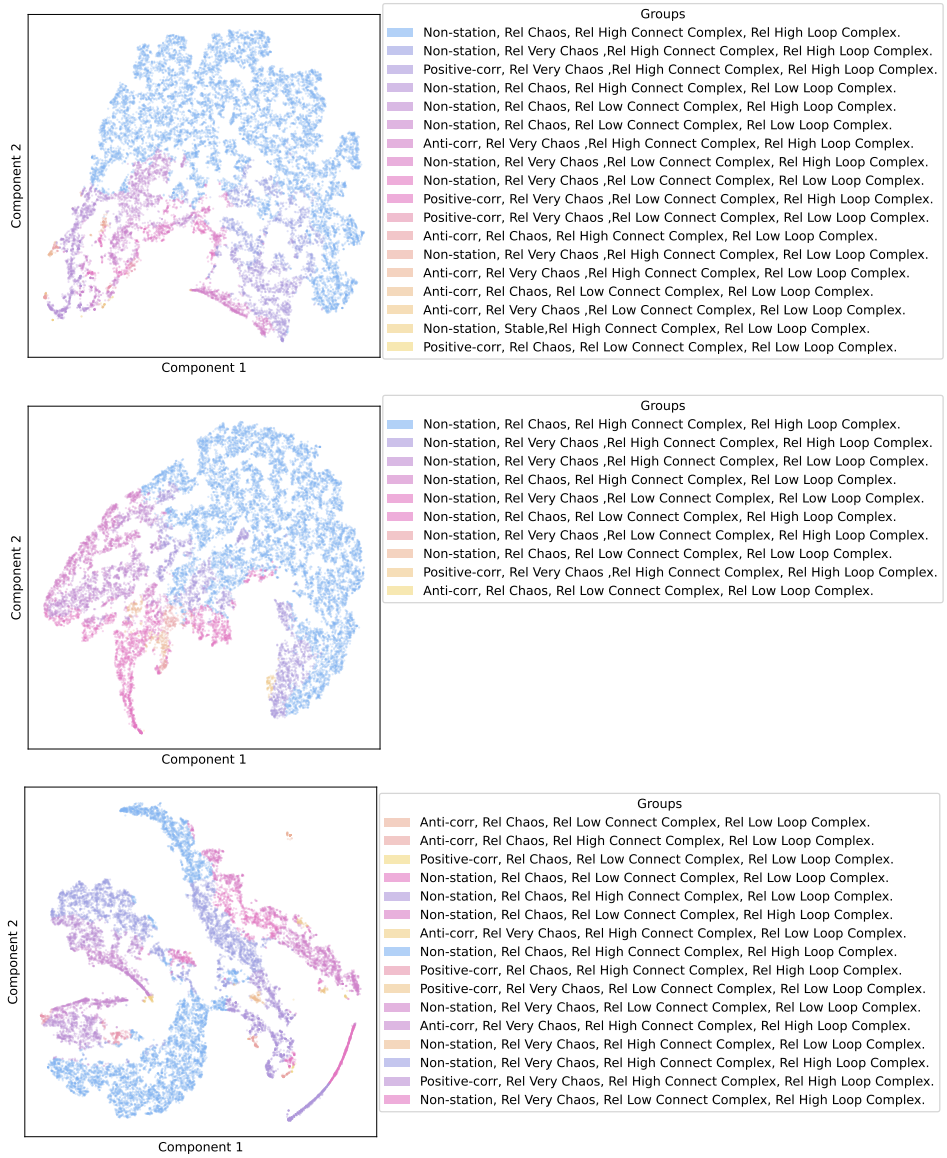

Figure 4: **T-SNE plot of the datasets, quantified by the proposed metrics from chaotic theory**. This figure mainly specified the exact group of time series with different chaotic attributes corresponding to plot presented in Figure 2 panel A.

E.2 Action Awareness Visualization

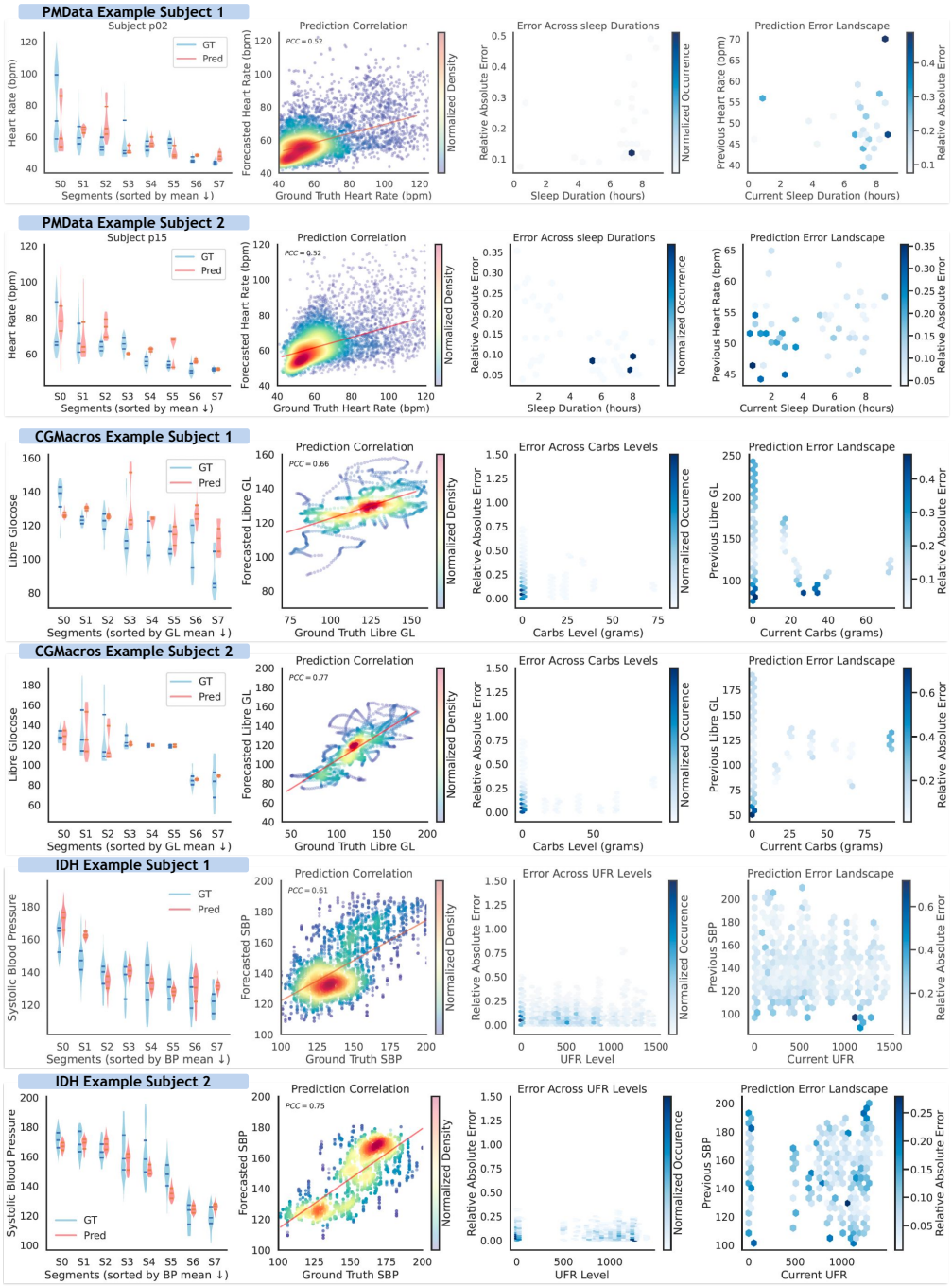

Figure 5: Supplementary results for Fig 3.D.: action/intervention analysis from example subjects.

E.3 VISUALIZING DYNAMICAL REASONING IN FORECASTING

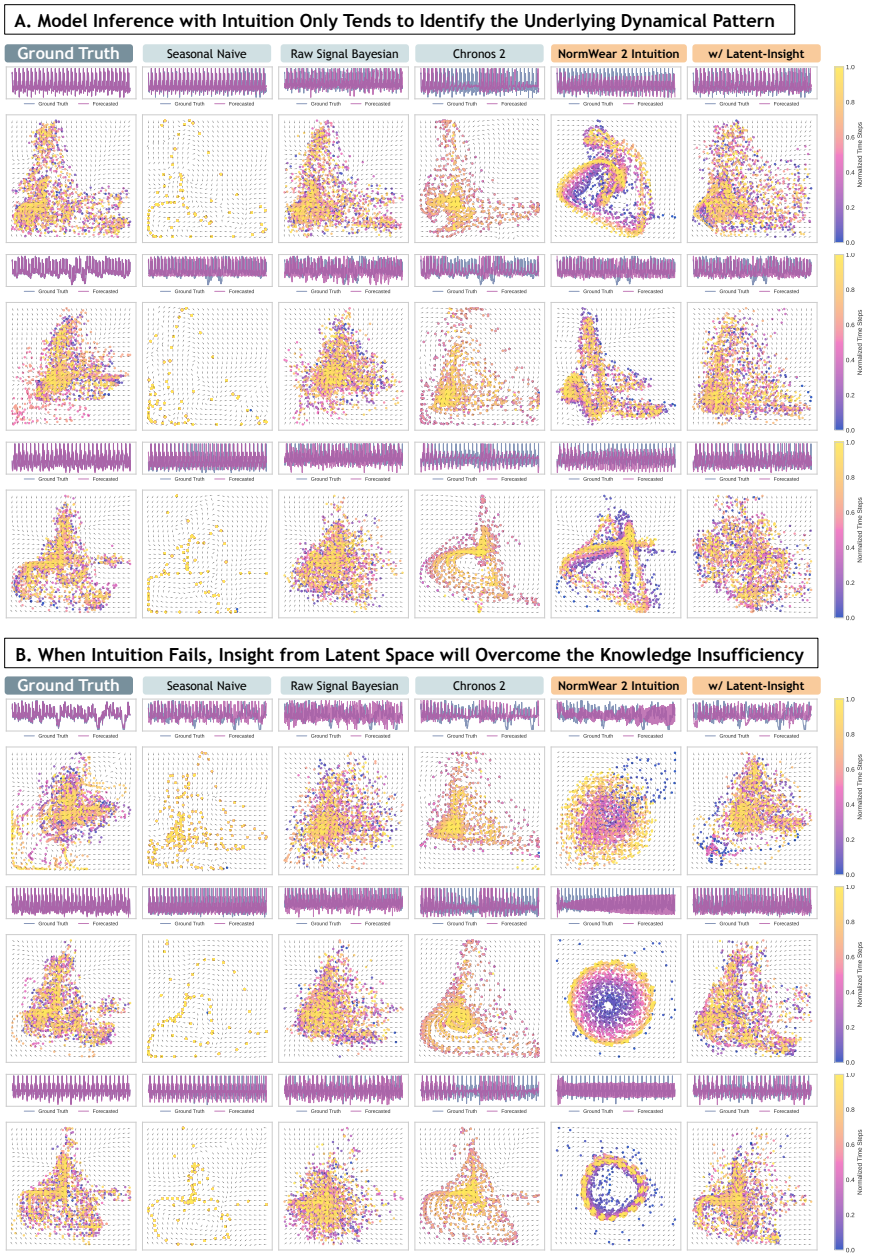

Figure 6: **Qualitative visualization of dynamical reasoning.** **(A)** Successful intuition-only forecasting reconstructs phase-space trajectories with better topology. **(B)** Failure cases where latent insight resolves ambiguity and improves forecasts. Color indicates normalized temporal progression.

To better understand the behaviors learned by NormWear 2 beyond quantitative metrics, we visualize both the pretrained latent representations and the forecasting dynamics in Figures 6 and 7. These visualizations provide qualitative evidence that the model captures meaningful temporal structure rather than relying solely on morphological based waveform matching. Figure 7 compares the temporal dynamics of representative physiological signals in raw signal space and pretrained latent space. Across ECG, EEG, and blood pressure examples, the latent representations preserve salient dynamical structures observed in the raw signals, including attractor geometry, temporal occupancy patterns, and state transition connectivity. This suggests that the pretrained latent space

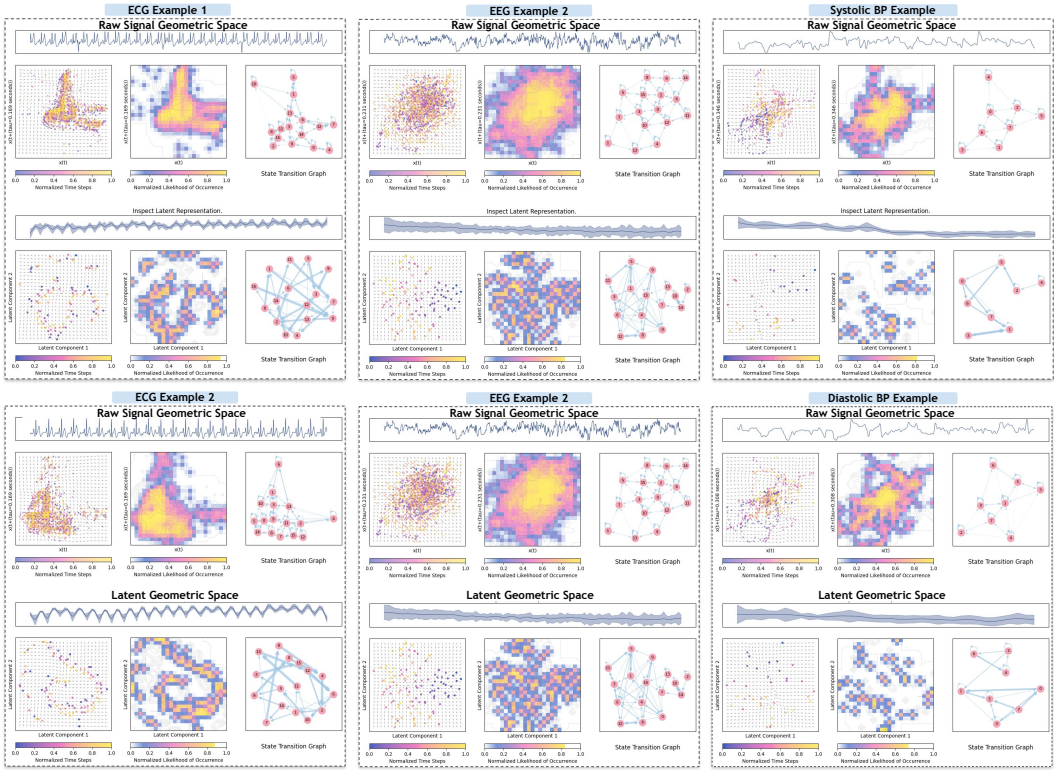

Figure 7: Comparison of raw-signal dynamics and pretrained latent dynamics across representative physiological signals, showing that latent representations preserve key temporal structures including attractor geometry, occupancy distribution, and transition connectivity.

retains the underlying temporal organization of physiological dynamics while compressing the raw observations into a more structured representation. Figure 6A illustrates cases where NormWear 2 successfully infers the underlying dynamical regime from partial observations using intuition alone. Although competing baselines often produce forecasts that match local waveform statistics, NormWear 2 generates forecasts whose reconstructed phase-space trajectories demonstrate more topological details, indicating that the model tends to identifying the underlying dynamics that could inform future trajectory progression as much as possible. Figure 6B highlights representative failure cases of intuition-only inference, where NormWear 2 did recognize a clear topological structure, but is insufficient to identify the future dynamics. In such ambiguous scenarios, incorporating latent-space insight resolves the uncertainty and produces forecasts that better recover the correct temporal evolution. This suggests that latent insight serves as a complementary mechanism to intuition-based forecasting, particularly when direct dynamical inference from partial observations is underdetermined. These visualizations indicate that NormWear 2 learns temporally structured latent representations and leverages them to perform forecasting through dynamical pattern reasoning.

E.4 Qualitative Visualization of Dynamical System Forecasting

Visualization of models' test time output is presented in Figure 8, 9, 10, and 11.

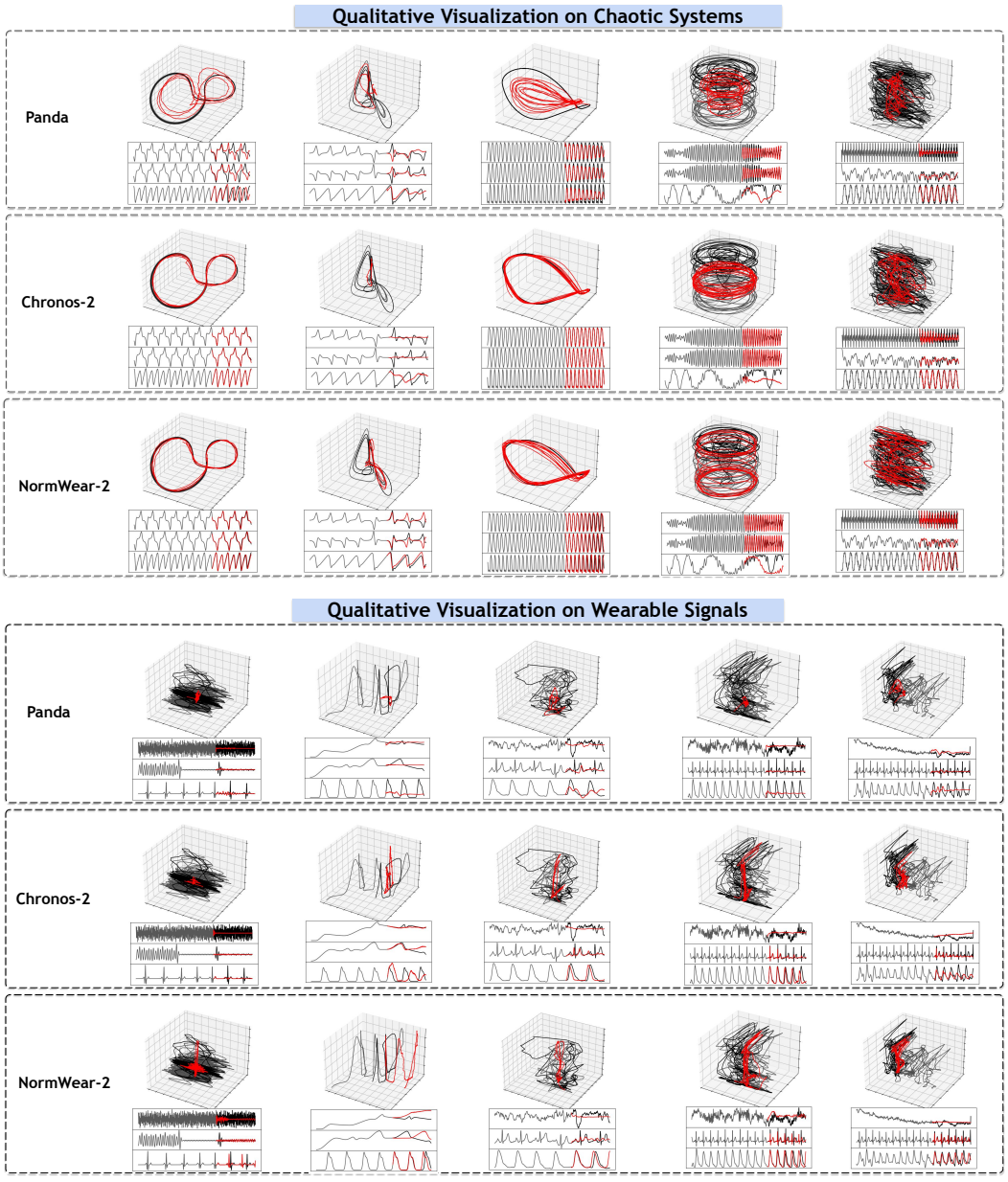

Figure 8: **Visualization of generation on time series randomly generated from test set**. Models in comparison are Panda (Lai et al., 2026) and Chronos (Ansari et al., 2024).

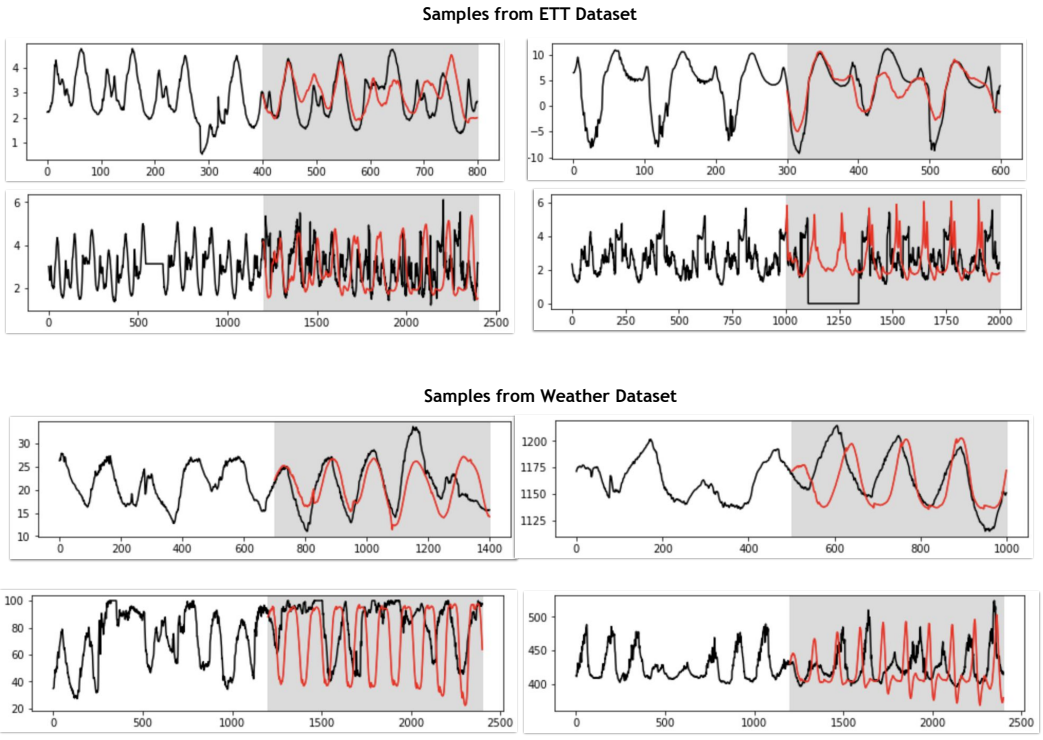

Figure 9: Visualization of generation on time series randomly generated from civil monitoring datasets proposed by Wu et al. (2021).

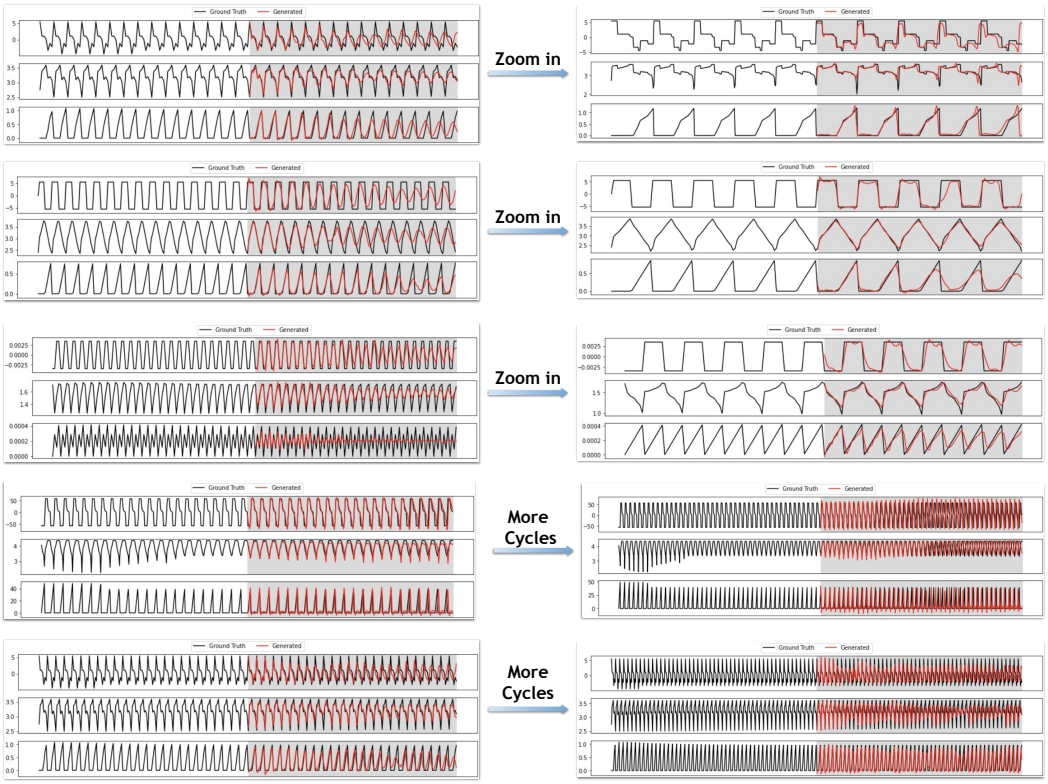

Figure 10: Visualization of generation on time series randomly generated from battery datasets proposed by Tan et al. (2025).

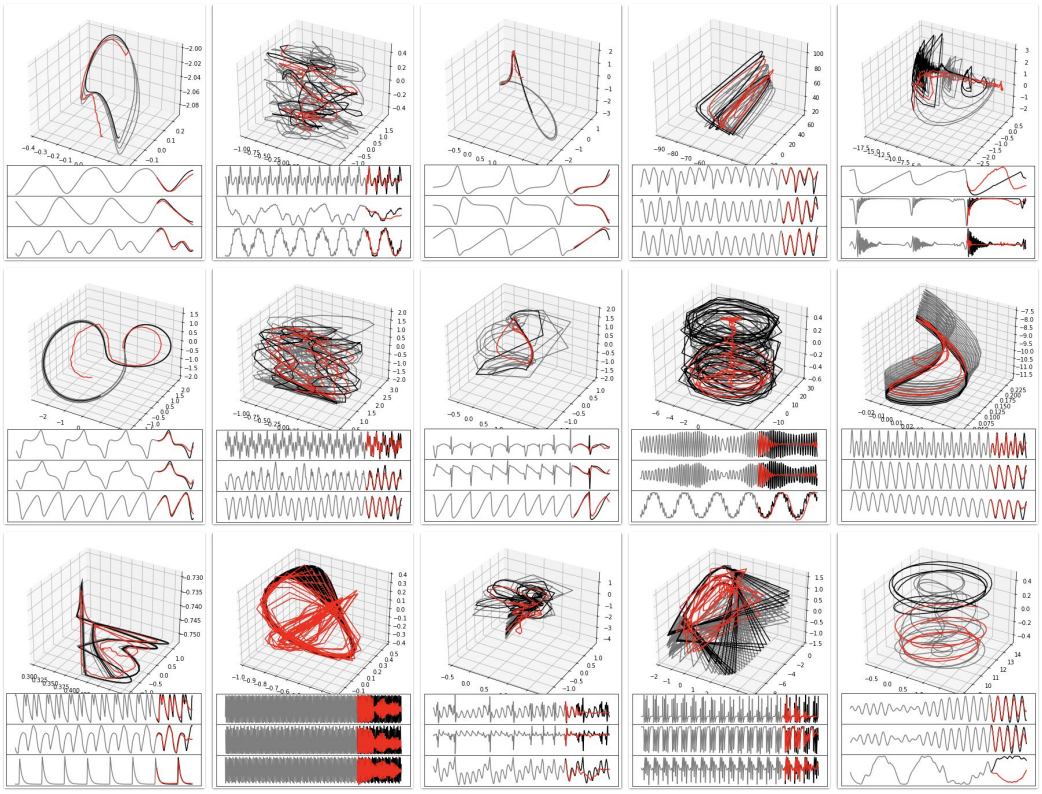

Figure 11: Visualization of generation on time series randomly generated from chaotic system datasets proposed by Gilpin (2021).