ECG IN MOTION: BENCHMARKING PHYSIOLOGICAL TRAJECTORY MODELING IN EMERGENCY CARE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in ECG representation learning have improved static tasks such as disease classification. Continuous monitoring in emergency care, however, requires models to predict how cardiac states evolve from continuous ECG history. Existing forecasting studies largely focus on short waveform segments and rely on pointwise errors that may obscure physiological fidelity and dynamic changes. We introduce ECG-WorldBench, a benchmark built on 60,127 history-future ECG waveform pairs from 8,259 patients, with patient-disjoint splits and 10-, 30-, and 60-second models conditioned on up to 300 seconds of history. It compares periodic, deterministic, and probabilistic models using complementary waveform, physiological, distributional, and semantic measures. We also propose PSLD, a physiologically structured latent diffusion model combining a physiology-supervised latent space with multiscale historical context. Experiments reveal a key trade-off: deterministic models minimize pointwise errors, whereas periodic continuation preserves recent beat morphology but poorly follows the realized future. Among learned methods, PSLD achieves the highest waveform correlation and R-peak F1 across all horizons and the strongest distributional and semantic agreement. By exposing these complementary capabilities, ECG-WorldBench provides a unified framework for moving ECG modeling beyond static recognition toward modeling physiological evolution in continuous emergency monitoring.

1 INTRODUCTION

Recent advances in ECG representation learning have improved cardiac classification and abnormality detection, enabling models to identify physiological features and pathological patterns in observed signals Chen et al. (2024); Jin et al. (2025); Ansari et al. (2025). Continuous monitoring in emergency care, however, requires not only determining whether an abnormality is present, but also understanding whether it is transient, persistent, or evolving, and anticipating possible subsequent changes to support early risk identification Sandau et al. (2017); Chen et al. (2023). Physiological changes may unfold across multiple beats and timescales, whereas isolated ECG segments provide only local snapshots of a patient's evolving condition. Modeling future ECG waveforms over multiple horizons from continuous history, and systematically evaluating how well they reflect physiological evolution, is therefore a central problem for continuous monitoring in emergency care.

Existing ECG generation and forecasting studies provide a foundation for this direction. Generative methods synthesize ECG waveforms with recognizable morphology, whereas forecasting methods infer subsequent segments from historical signals Lai et al. (2025); Wang et al. (2026b); Prakarsha & Sharma (2022); Tang et al. (2025). However, existing studies largely focus on short waveform segments and do not fully address the long-history, multi-horizon setting required for continuous emergency monitoring. Because ECG is quasi-periodic, strong short-horizon performance may primarily reflect the continuation of local beat patterns rather than effective modeling of subsequent physiological changes (Kundu et al. (2000); Berkaya et al. (2018)). Moreover, pointwise objectives may produce overly smooth waveforms, reducing prediction error at the expense of physiological structure. Evaluation must therefore go beyond pointwise error and jointly assess waveform accuracy, physiological fidelity, and representation-level consistency across forecast horizons.

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.

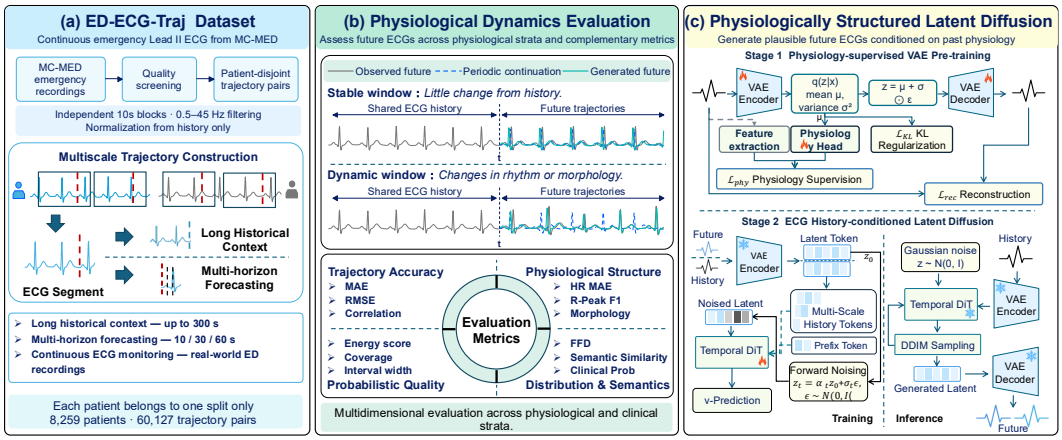


Figure 1: Overview of ECG-WorldBench. (a) ED-ECG-Traj construction from continuous emergency ECG for long-context, multi-horizon forecasting. (b) Multidimensional evaluation across physiological and clinical strata. (c) PSLD, combining physiology-supervised VAE pretraining with history-conditioned latent diffusion.

Against this background, modeling physiological evolution from continuous ECG in emergency care remains challenging at the data, evaluation, and modeling levels:

- **Data level:** Continuous monitoring records must preserve temporal relationships across segments. Each example should pair a long ECG history with subsequent waveforms at multiple forecast horizons, while data splits should remain patient-disjoint.
- **Evaluation level:** A single waveform error or population-level realism metric cannot adequately distinguish the preservation of local patterns from the modeling of subsequent changes. Evaluation should therefore jointly assess waveform accuracy, physiological structure, and semantic consistency across multi-scale horizons.
- **Modeling level:** Long ECG histories contain both local beat structure and physiological changes spanning multiple beats. Models must integrate historical information across multiple timescales, preserve patient-specific physiological structure, and generate plausible future waveforms.

To address these challenges, we introduce ECG-WorldBench, a benchmark for modeling future ECG waveforms from continuous histories in emergency care, as shown in Figure 1. ECG-WorldBench brings together a continuous ECG dataset, which is curated from MC-MED (Kansal et al. (2025)), a multidimensional evaluation protocol, and a generative model to support systematic and reproducible study of physiological evolution across multiple forecast horizons. Our contributions are summarized as follows:

- **Emergency ECG waveform dataset.** We construct ED-ECG-Traj from continuous Lead II ECG in MC-MED, comprising 60,127 history-future ECG waveform pairs from 8,259 patients. With patient-disjoint splits and strict temporal boundaries, each example pairs up to 300 seconds of ECG history with future waveforms at 10-, 30-, and 60-second horizons.
- **Physiological dynamics evaluation.** We establish a multidimensional evaluation framework that uses patient-specific periodic continuation as a strong baseline and jointly assesses waveform accuracy, physiological structure, and semantic consistency across horizons. Dynamics stratification and statistical analysis reveal discrepancies between pointwise accuracy and physiological fidelity, while probabilistic metrics provide additional evaluation for generative models.
- **Physiologically structured waveform modeling.** We propose PSLD, a conditional latent diffusion model that combines a physiology-supervised latent space with multiscale ECG-history encoding, a clean recent-waveform prefix, and causal rhythm-phase conditioning.

2 Related Work

2.1 ECG Representation Learning

ECG representation learning aims to extract informative features from cardiac signals for tasks such as rhythm classification, abnormality detection, and clinical outcome prediction. Beyond supervised learning, self-supervised approaches use objectives such as contrastive learning and masked signal reconstruction to learn representations from unlabeled recordings (Na et al. (2024); Yuan et al. (2025); Wang et al. (2026a); She et al. (2026)). Multimodal approaches further align ECG signals with clinical text, connecting waveform features with semantic information (Liu et al. (2024); Chen et al. (2025; 2026a)). These developments provide useful representations for downstream tasks, although performance on recognition tasks does not directly establish the ability to model subsequent physiological evolution. Building on these advances, we learn a decodable, physiologically structured representation for conditional ECG waveform generation, while frozen ECG encoders assess representational and semantic consistency beyond waveform and physiological metrics.

2.2 ECG Generation and Forecasting

ECG generation methods commonly use generative adversarial networks or diffusion models conditioned on diagnostic labels or clinical text (Kranz et al. (2025); Kim & Jung (2026); Wang et al. (2026c)), but waveform realism does not ensure consistency with a patient's ECG history. ECG forecasting methods use historical signals to predict subsequent waveforms (Tang et al. (2025)), which focus on short historical and forecast segments and primarily optimize pointwise errors, with limited attention to physiological evolution and structural preservation across timescales. Our work targets continuous monitoring in emergency care by studying multi-horizon ECG waveform modeling from long historical contexts and enhancing consistency with a patient's preceding physiological state through a physiologically structured latent space.

2.3 ECG Datasets and Benchmarks

ECG datasets support diverse research settings, ranging from diagnostic recordings with rhythm and disease annotations to continuous physiological monitoring (Liu et al. (2018); Zheng et al. (2022); Wagner et al. (2020); Gow et al. (2023); Al-Masud et al. (2026)). Among these resources, MC-MED (Kansal et al. (2025)) provides the emergency-care recordings used to construct ED-ECG-Traj. Existing benchmarks organize ECG learning into tasks such as classification, detection, forecasting, and generation. ECG-Benchmark (Tang et al. (2025)), in particular, evaluates forecasting alongside other tasks and introduces feature Fréchet distance (FFD) to complement conventional signal-level metrics. ECG-WM studies action-conditioned post-intervention ECG simulation using ODE-guided latent diffusion Chen et al. (2026b). In contrast, ECG-WorldBench studies observation-conditioned future waveforms from continuous emergency ECG. It uses patient-disjoint, multi-horizon pairs, strict observation boundaries, and complementary evaluation across stable and changing conditions.

3 Continuous ECG Benchmark for Emergency Care

ECG-WorldBench combines ED-ECG-Traj, multidimensional evaluation, and a conditional generative model to study future ECG evolution from continuous emergency-care recordings. Given an ECG history preceding modeling origin t, the task is to generate a future waveform of length H:

$$\hat{x}^{(k)}_{t:t+H} = f_\theta(x_{t-L:t}, \epsilon_k)\,, \qquad H \in \{10, 30, 60\}\,\text{s}.$$

Here, $L \leq 300$ seconds denotes the method-specific history length. Deterministic methods produce one future waveform, whereas probabilistic methods generate multiple samples through ϵ_k. All methods use the same observation cutoff and test windows.

3.1 ED-ECG-Traj Dataset Construction

We construct ED-ECG-Traj from continuous Lead II ECG recordings in MC-MED to study multi-horizon ECG waveform generation in emergency care. The original waveforms are sampled at 500

(a) Cohort statistics

Statistic	Train	Val	Test	Overall
Patients	7,436	407	416	8,259
ED encounters	9,531	509	531	10,571
Age (years)	62 [46,76]	61 [45,75]	60 [43,74]	62 [46,76]
Female (%)	51.2	51.8	45.2	50.9
Trajectory pairs	54,162	3,011	2,954	60,127
Pairs/patient	10 [4,10]	10 [5,10]	9 [4,10]	10 [4,10]
ECG coverage (h)	3,471.72	191.86	186.91	3,850.49

(b) Clinical weak labels

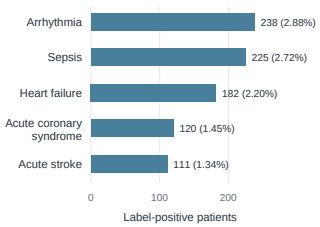


Figure 2: ED-ECG-Traj cohort statistics and clinical label distribution.

Hz. For each forecast origin, we use the preceding 300 seconds of continuous ECG as historical input and pair it with future waveforms of 10, 30, and 60 seconds, forming a unified set of history-future examples. Constructing such examples from continuous recordings requires ensuring signal quality while preventing future information leakage. We first exclude incomplete windows and those severely affected by artifacts through technical quality screening, with all data-driven thresholds calibrated exclusively on the training set. Waveforms are then processed in independent 10-second blocks using a zero-phase 0.5–45 Hz band-pass filter, preventing future signals from influencing historical inputs through filtering. Normalization parameters are likewise estimated solely from the preceding 300-second history and applied consistently to both historical and future waveforms. All methods are evaluated in this common preprocessing space, with waveform metrics computed directly against the preprocessed observed future rather than reconstructions from a latent model.

After establishing valid temporal boundaries, we further control data independence and unequal patient contributions through patient-level splitting and sampling. Training, validation, and test sets follow a 90:5:5 patient-level split, with no patient overlap across partitions. To prevent patients with long recordings from being overrepresented, we apply deterministic within-patient sampling: routine windows are selected uniformly along each patient timeline and are typically capped at 10 per patient, while all eligible recorded-action-associated windows are retained. ED-ECG-Traj ultimately contains 60,127 history–future waveform pairs from 8,259 patients. For each forecast horizon, all methods are evaluated on the same fixed set of eligible test examples. Figure 2 summarizes the dataset splits, cohort statistics, and distribution of clinical weak labels.

3.2 Physiological Dynamics Evaluation

Evaluating future ECG waveforms requires assessing both their agreement with observed outcomes and the preservation of physiologically meaningful beat structure. Pointwise error alone is insufficient: lower error does not necessarily indicate more accurate rhythm or beat morphology, while periodic continuation may remain competitive under stable physiological conditions. We therefore evaluate models by combining complementary waveform and physiological metrics with periodic baselines and dynamics stratification.

Evaluation protocol and periodic baselines. We evaluate all methods on the same test windows and report results separately across forecast horizons. Each method uses only information available before the forecast origin. The evaluation includes two training-free periodic baselines: Last-10s Repeat directly repeats the most recent 10-second waveform segment, whereas Phase-Aligned Median Beat extends a patient-specific beat template using the estimated rhythm and boundary phase. Together, these baselines test whether learned models improve beyond the simple continuation of recent waveform structure.

Complementary evaluation dimensions. We evaluate generated future ECG waveforms along four complementary dimensions. Detailed metric definitions are provided in Appendix A.3:

- **Waveform accuracy.** MAE, RMSE, and correlation measure pointwise error and overall agreement between generated and observed future waveforms.
- **Physiological structure.** Heart Rate (HR) MAE, R-peak F1, and beat-aligned morphology distance assess the preservation of cardiac rhythm, beat timing, and waveform morphology.

- **Probabilistic quality.** For methods that generate multiple future waveforms, we report the energy score together with prediction-interval coverage and width.
- **Distributional and semantic consistency.** Using the frozen ECG-Benchmark (Tang et al. (2025)) feature extractor, ECGFounder (Li et al. (2025a)), and CLEF-Small (Shu et al. (2025)), we assess distributional distance, representation similarity, and ECGFounder classification-probability differences between generated and observed future waveforms.

Dynamics stratification. Aggregate results can obscure differences in model performance between physiologically stable and changing conditions. We therefore partition evaluation windows into stable and dynamic strata according to predefined ECG-based criteria. These labels are used exclusively for evaluation and are not provided as model inputs. The stratification enables separate assessment of physiological structure preservation and trajectory modeling under observed changes. Detailed stratification rules and thresholds are provided in Appendix A.2.

3.3 Physiologically Structured Latent Diffusion

As illustrated in Figure 1, we propose PSLD, a two-stage generative model that combines physiology-supervised representation learning with ECG history-conditioned latent diffusion.

Physiology-supervised VAE pre-training. In the first stage, we train a variational autoencoder (VAE) to encode ECG waveforms into compact latent sequences and reconstruct the original signals. A prediction head further uses the full latent representation to estimate waveform-derived descriptors of cardiac rhythm and beat morphology. The training objective combines reconstruction loss, physiological supervision, and KL regularization:

$$\mathcal{L}_{\mathrm{VAE}} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}} + \beta\mathcal{L}_{\mathrm{KL}}. \tag{1}$$

Physiological supervision encourages the latent space to preserve rhythm- and morphology-related information. After pre-training, the encoder and decoder are frozen during the second stage.

ECG history-conditioned latent diffusion. In the second stage, we encode the observed ECG history into multiscale history features and local features near the forecasting boundary, capturing both long-range context and recent changes. These features, together with rhythm and phase information derived from the observed history, condition a temporal diffusion transformer. During training, Gaussian noise is added to the latent representation of the observed future waveform, and the transformer learns to denoise it using a velocity-prediction objective:

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}\left[\left\|v_\theta(z_\tau, \tau, c_{\mathrm{hist}}) - v_\tau\right\|_2^2\right], \tag{2}$$

where z_τ denotes the noised future latent representation, τ is the diffusion timestep, and c_{hist} contains only information derived from the observed history. At inference, DDIM sampling generates future latent representations from independent Gaussian noise, which are then mapped to ECG waveforms by the frozen VAE decoder. Repeated sampling produces multiple possible future waveforms conditioned on the same history. Further details are provided in Appendix A.5.

4 Experiments

We evaluate ECG-WorldBench across multiple future horizons and physiological dynamics. Our experiments address three questions: (1) how periodic, deterministic, and probabilistic methods compare in waveform accuracy, physiological fidelity, and semantic consistency; (2) whether learned models capture future changes beyond the continuation of recent periodic patterns; and (3) how their performance varies between stable and dynamically changing ECG windows.

Data and evaluation protocol. We use the patient-disjoint training, validation, and test splits of ED-ECG-Traj described in Section 3.1 and evaluate future ECG waveforms over 10-, 30-, and 60-second horizons. All learned methods use the 300 seconds of ECG history preceding the forecasting origin and are compared on the same test windows using the evaluation framework introduced in Section 3.2. We also provide history-relative change analysis and patient-level paired significance tests in Appendix A.12 and Appendix A.13.

Table 1: Waveform accuracy and physiological structure comparison. Bold denotes the best.

	Waveform accuracy			Physiological structure		
Method	MAE ↓	RMSE ↓	Corr. ↑	HR MAE ↓	R-peak F1 ↑	Morph. ↓
10s horizon						
Last-10s Repeat	1.710	2.508	0.008	**5.712**	0.169	**0.182**
Phase-Aligned Median Beat	1.405	2.205	0.129	22.531	0.353	0.437
PatchTST	1.257	1.866	0.014	30.399	0.177	1.342
PSSM	**1.154**	**1.767**	0.106	29.839	0.163	1.353
Latent-NsDiff	1.397	1.983	0.004	30.255	0.174	1.307
Latent-D3U	1.408	2.001	0.031	30.176	0.178	1.312
PSLD (Ours)	1.346	2.130	**0.186**	7.198	**0.422**	0.271
30s horizon						
Last-10s Repeat	1.715	2.521	0.006	**5.296**	0.164	**0.164**
Phase-Aligned Median Beat	1.509	2.318	**0.057**	22.754	**0.250**	0.420
PatchTST	1.225	1.838	0.007	29.805	0.171	1.340
PSSM	**1.172**	**1.789**	0.019	30.641	0.175	1.333
Latent-NsDiff	1.401	1.989	0.001	29.326	0.169	1.260
Latent-D3U	1.436	2.033	0.004	29.043	0.169	1.265
PSLD (Ours)	1.422	2.154	0.052	6.869	0.231	0.287
60s horizon						
Last-10s Repeat	1.718	2.528	0.005	**5.368**	0.164	**0.161**
Phase-Aligned Median Beat	1.549	2.358	**0.031**	22.916	**0.212**	0.418
PatchTST	1.227	1.846	0.004	29.212	0.171	1.338
PSSM	**1.172**	**1.793**	0.004	30.657	0.172	1.330
Latent-NsDiff	1.401	1.992	0.001	28.615	0.170	1.194
Latent-D3U	1.447	2.046	0.000	28.800	0.169	1.219
PSLD (Ours)	1.468	2.206	0.023	5.994	0.190	0.254

Baselines. We compare against six baselines across three categories. **Training-free periodic baselines** include Last-10s Repeat and Phase-Aligned Median Beat. The former repeats the most recent 10-second waveform, whereas the latter constructs a median beat template from recent history and extends it according to the estimated rhythm and boundary phase. **Deterministic learned baselines** include the time-series forecasting model PatchTST (Nie et al. (2022)) and the ECG-specific PSSM (Tang et al. (2025)) adapted with a forecasting head. **Probabilistic generative baselines** include Latent-NsDiff (Ye et al. (2025)) and Latent-D3U (Li et al. (2025b)), which generate multiple future waveforms conditioned on the observed history. Further details are provided in Appendix A.6.

Implementation details. All baselines use the same patient-disjoint data splits, preprocessed Lead II waveforms, and evaluation protocol. To reduce the computational cost of directly generating long, high-frequency ECG waveforms, Latent-NsDiff and Latent-D3U are adapted to the same frozen Stage-1 VAE latent space as PSLD and use the same latent representation and decoder. All experiments are conducted on four NVIDIA GeForce RTX 4090 GPUs.

4.1 Waveform Accuracy and Physiological Fidelity

Table 1 compares waveform accuracy and physiological fidelity across the three future horizons. PSSM achieves the lowest MAE and RMSE at all horizons, whereas PSLD consistently obtains the highest waveform correlation and R-peak F1 among learned methods. PSLD also substantially reduces HR error and morphology distance relative to the other learned baselines.

10s horizon. PSLD achieves the highest correlation (0.186) and R-peak F1 (0.422), with the lowest HR MAE (7.198) and morphology distance (0.271) among learned methods. It outperforms Phase-Aligned Median Beat on all metrics. PSSM obtains the lowest MAE and RMSE (1.154 and 1.767), while Last-10s Repeat achieves the lowest overall HR error and morphology distance.

30s horizon. PSLD obtains the highest correlation (0.052) and R-peak F1 (0.231) among learned methods, while substantially reducing HR error and morphology distance relative to the other learned baselines. It outperforms the Phase-Aligned Median Beat on most metrics, although the periodic baseline achieves higher correlation (0.057) and R-peak F1 (0.250). PSSM retains the lowest pointwise errors, while Last-10s Repeat is the lowest overall HR error and morphology distance.

Table 2: Consistency evaluation of learned methods across horizons. Bold marks the best result; interval width is unranked. Coverage and width are computed from empirical 5th-95th percentile intervals using $K = 8$ generated waveforms and are interpreted jointly.

	Probabilistic quality			Distribution	ECGFounder		CLEF
Method	Energy ↓	Cov. (5-95%)	Width	FFD ↓	Cos. ↑	Prob. MAE ↓	Cos. ↑
10s horizon							
PatchTST	–	–	–	5.633	0.411	0.1098	0.729
PSSM	–	–	–	5.965	0.399	0.1053	0.667
Latent-NsDiff	1.447	0.5887	2.149	5.609	0.418	0.1125	0.732
Latent-D3U	1.427	0.6225	2.250	5.564	0.416	0.1119	0.735
PSLD (Ours)	**1.287**	**0.6434**	2.671	**0.157**	**0.864**	**0.0351**	**0.950**
30s horizon							
PatchTST	–	–	–	6.534	0.405	0.1081	0.719
PSSM	–	–	–	6.855	0.396	0.1088	0.685
Latent-NsDiff	1.452	0.5889	2.159	6.616	0.419	0.1125	0.733
Latent-D3U	1.448	**0.6241**	2.300	6.689	0.415	0.1120	0.732
PSLD (Ours)	**1.375**	0.5966	2.626	**0.835**	**0.797**	**0.0524**	**0.933**
60s horizon							
PatchTST	–	–	–	6.900	0.405	0.1084	0.721
PSSM	–	–	–	7.196	0.396	0.1106	0.696
Latent-NsDiff	1.457	0.5880	2.150	7.003	0.419	0.1128	0.732
Latent-D3U	1.452	**0.6291**	2.336	7.168	0.414	0.1123	0.731
PSLD (Ours)	**1.397**	0.6022	2.735	**0.695**	**0.853**	**0.0411**	**0.942**

60s horizon. PSLD obtains the highest correlation (0.023) and R-peak F1 (0.190), together with the lowest HR MAE (5.994) and morphology distance (0.254), among learned methods. It outperforms Phase-Aligned Median Beat on four of the six metrics, while the periodic baseline achieves higher correlation (0.031) and R-peak F1 (0.212). PSSM retains the lowest MAE and RMSE, whereas Last-10s Repeat achieves the lowest overall HR error and morphology distance.

4.2 Probabilistic Quality and Semantic Consistency

Table 2 reports probabilistic quality, feature-space distributional agreement, and semantic consistency for the learned methods. For probabilistic generative methods, we report the energy score and the coverage and width of empirical 5th-95th percentile intervals estimated from $K = 8$ trajectories. We interpret empirical coverage jointly with interval width under the common sampling budget of $K = 8$; finite-sample effects are discussed in Appendix A.16.2. We additionally evaluate FFD, ECGFounder and CLEF-Small representation similarities, and ECGFounder classification-probability MAE.

10s horizon. PSLD achieves the lowest energy score (1.287) and the highest empirical coverage (0.6434) among the probabilistic methods, although it also produces the widest interval (2.671). PSLD further achieves the lowest FFD (0.157) and the strongest three semantic metrics.

30s horizon. PSLD maintains the lowest energy score (1.375) and achieves the lowest FFD (0.835), together with the strongest ECGFounder and CLEF-Small consistency. Its empirical coverage is 0.5966, compared with 0.6241 for Latent-D3U, despite having a wider interval.

60s horizon. PSLD again achieves the lowest energy score (1.397) and FFD (0.695) and retains the strongest ECGFounder and CLEF-Small consistency. Its empirical coverage is 0.6022, compared with 0.6291 for Latent-D3U, while its interval is wider (2.735 versus 2.336). Overall, PSLD consistently improves the energy score and representation-level metrics, whereas the coverage-width comparisons show no uniform advantage.

4.3 Performance across Physiological Dynamics

To examine whether the aggregate findings remain consistent across different ECG dynamics, we evaluate all methods separately within the stable and dynamic strata defined in Section 3.2, focusing on the 10-second horizon. Figure 3 reports results for 2,005 stable windows from 379 patients and

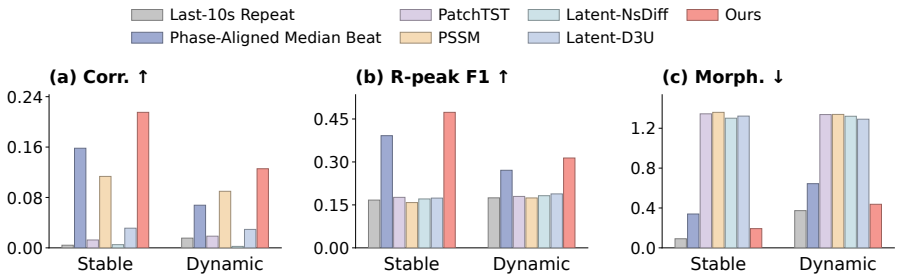


Figure 3: Waveform correlation, R-peak F1, and morphology distance across stable and dynamic ECG windows at the 10-second horizon.

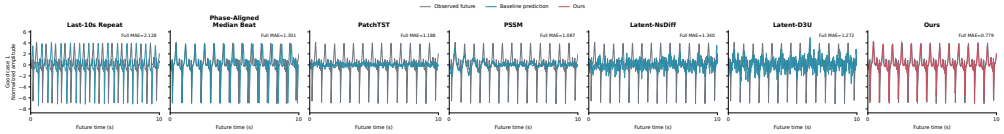


Figure 4: 10s case visualization. Gray: observed ECG; blue: baseline predictions; red: ours.

949 dynamic windows from 288 patients. PSLD achieves the highest waveform correlation and R-peak F1 in both strata, together with the lowest morphology distance among learned methods. Its correlation, R-peak F1, and morphology distance are 0.215, 0.473, and 0.193 in stable windows, compared with 0.126, 0.314, and 0.437 in dynamic windows. PSLD outperforms Phase-Aligned Median Beat on all three metrics in both strata, whereas Last-10s Repeat retains the lowest overall morphology distance. PSLD nevertheless performs worse in the dynamic stratum than in the stable stratum across all three metrics, indicating that the threshold-defined dynamic windows remain more challenging to model. More results and analyses are provided in Appendix A.9.

5 Analysis and Discussion

5.1 Qualitative comparison.

Figure 4 compares seven methods on a 10-second forecasting example. PatchTST shows substantial amplitude attenuation, while PSSM reconstructs some early beats but weakens later ones. The periodic baselines preserve the main beat complexes but exhibit temporal misalignment, whereas the latent diffusion baselines produce irregular fluctuations with weaker correspondence to the observed future. PSLD follows the main complexes more closely, although local timing and amplitude errors remain. Its MAE is 0.779, compared with 1.087 for PSSM and 1.188 for PatchTST. Additional favorable and failure cases are provided in Appendix A.18.

5.2 Strengths and Limitations of Modeling Paradigms

ECG-WorldBench reveals distinct strengths across periodic continuation, deterministic prediction, and probabilistic generation.

Periodic continuation methods. Last-10s Repeat achieves the lowest morphology distance across all horizons, highlighting the strong persistence of recent beat structure, while its near-zero waveform correlation indicates limited temporal correspondence with the observed future. Phase-Aligned Median Beat improves waveform correlation and R-peak F1 by explicitly modeling rhythm and boundary phase, at the cost of higher HR and morphology errors. Periodic continuation therefore provides a strong reference for rhythm and morphology preservation.

Deterministic prediction methods. PSSM achieves the lowest MAE and RMSE across all horizons, with PatchTST showing similar advantages in pointwise accuracy. Their larger HR and morphology errors and weaker semantic consistency, however, indicate that pointwise optimization captures a different aspect of future waveform modeling from physiological and representation-level preservation.

Probabilistic generation methods. Among learned methods, PSLD achieves the lowest energy score and FFD across all horizons and the strongest ECGFounder and CLEF-Small consistency. It also provides stronger physiological fidelity than the other learned baselines, although recent-beat morphology is preserved more effectively by Last-10s Repeat. These results position PSLD as a balanced probabilistic model across temporal, physiological, distributional, and semantic dimensions.

5.3 Effects of Future Horizons

As the horizon increases from 10 to 60 seconds, waveform correlation decreases substantially for most methods, indicating that temporal agreement with the observed future becomes more difficult to maintain. R-peak F1 also decreases for the periodic methods and PSLD, while the other learned baselines show smaller changes. In contrast, pointwise errors do not follow a consistent trend: MAE and RMSE increase for the periodic methods and PSLD but remain relatively stable or decrease slightly for the deterministic methods and other latent generative baselines. The future horizon therefore has a more consistent effect on waveform correlation than on pointwise error.

Other metrics exhibit different horizon-dependent trends. HR error and morphology distance do not increase monotonically, and some methods achieve lower values over longer horizons; however, improvements in these window-level metrics do not necessarily indicate better long-horizon waveform modeling. Energy scores increase with the future horizon, and prediction intervals are generally wider at 60s than at 10s, while coverage shows no consistent trend. FFD increases with the horizon for the other learned methods, whereas PSLD shows non-monotonic changes in FFD and semantic consistency while retaining the best performance across all three horizons. Overall, longer future horizons most consistently reduce temporal agreement, while their effects on physiological, distributional, and semantic metrics are more complex. In addition to the future horizon, Appendix A.11 examines the effect of observed context length through history-length ablations.

5.4 Agreement and Disagreement across Evaluation Metrics

Different metrics yield different method rankings. PSSM achieves the lowest MAE and RMSE across all horizons, whereas Last-10s Repeat best preserves recent beat morphology despite its near-zero waveform correlation. PSLD provides more balanced performance across waveform correlation, physiological structure, and representation-level consistency. These differences confirm that pointwise accuracy, temporal agreement, and physiological fidelity capture complementary capabilities. Probabilistic metrics exhibit a similar trade-off. PSLD achieves the lowest energy scores, whereas its wider intervals do not consistently yield higher empirical coverage than Latent-D3U. Moreover, even the highest waveform correlations remain low, highlighting the difficulty of precisely predicting the temporal evolution of future ECGs. We report coverage jointly with interval width, with finite-sample considerations detailed in Appendix A.16.2. PSLD's consistent advantages across ECGFounder and CLEF-Small further support its representation-level consistency.

6 Conclusion

We introduce ECG-WorldBench, a unified benchmark for modeling subsequent ECG trajectories conditioned on continuous history in emergency care. It comprises three key contributions: ED-ECG-Traj, a dataset of 60,127 history–future waveform pairs from 8,259 patients, with patient-disjoint splits, up to 300 seconds of historical context, and 10-, 30-, and 60-second future horizons; a multidimensional evaluation protocol that jointly assesses waveform accuracy, physiological structure, probabilistic quality, and representation-level consistency across stable and dynamic conditions; and PSLD, a physiologically structured latent diffusion model that combines physiology-supervised representation learning with multiscale historical and rhythm-phase conditioning. Experiments reveal complementary strengths across modeling paradigms: deterministic methods minimize pointwise errors, periodic continuation best preserves recent beat morphology, and PSLD achieves the strongest overall balance among learned methods, with the highest waveform correlation and R-peak F1 together with the best distributional and semantic consistency. These findings demonstrate that waveform error alone is insufficient for evaluating future ECG modeling and establish ECG-WorldBench as a reproducible foundation for studying history-conditioned ECG evolution in continuous emergency monitoring.

REFERENCES

MA Al-Masud, Juan Miguel Lopez Alcaraz, and Nils Strodthoff. Benchmarking ecg fms: A reality check across clinical tasks. In *International Conference on Learning Representations*, volume 2026, pp. 146689–146715, 2026.

Mohammed Yusuf Ansari, Mohammed Yaqoob, Mohammed Ishaq, Eduardo Feo Flushing, Iffa Afsa Changaai Mangalote, Sarada Prasad Dakua, Omar Aboumarzouk, Raffaella Righetti, and Marwa Qaraqe. A survey of transformers and large language models for ecg diagnosis: advances, challenges, and future directions. *Artificial Intelligence Review*, 58(9):261, 2025.

Selcan Kaplan Berkaya, Alper Kursat Uysal, Efnan Sora Gunal, Semih Ergin, Serkan Gunal, and M Bilginer Gulmezoglu. A survey on ecg analysis. *Biomedical Signal Processing and Control*, 43:216–235, 2018.

Emma Chen, Aman Kansal, Julie Chen, Boyang Tom Jin, Julia Reisler, David E Kim, and Pranav Rajpurkar. Multimodal clinical benchmark for emergency care (mc-bec): A comprehensive benchmark for evaluating foundation models in emergency medicine. *Advances in Neural Information Processing Systems*, 36:45794–45811, 2023.

Jian Chen, Yuzhu Hu, Lalit Garg, Thippa Reddy Gadekallu, Gautam Srivastava, and Wei Wang. Graph-enhanced low-resource ecg representation learning for emotion recognition based on wearable internet of things. *IEEE Internet of Things Journal*, 11(24):39056–39068, 2024.

Jian Chen, Xiaoru Dong, Wei Wang, Shaorui Zhou, Lequan Yu, and Xiping Hu. Deri: Cross-modal ecg representation learning with deep ecg-report interaction. In *IJCAI*, pp. 4824–4832, 2025.

Jian Chen, Yipeng Du, Wenhao Yuan, Shuai Wang, Jinfeng Xu, Zewei Liu, Running Zhao, and Edith Cheuk-Han Ngai. Sgera: Stein-guided ecg-report alignment for ecg representation learning. In *Forty-third International Conference on Machine Learning*, 2026a.

Zhikang Chen, Yue Wang, Sen Cui, Yu Zhang, Changshui Zhang, Tianling Ren, and Tingting Zhu. Ecg-wm: A physiology-informed ecg world model for clinical intervention simulation. *arXiv preprint arXiv:2605.17580*, 2026b.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset. *Type: dataset*, 2023.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ecg words and sentences via pre-training ecg language model. In *International conference on learning representations*, volume 2025, pp. 8207–8227, 2025.

Aman Kansal, Emma Chen, Boyang Tom Jin, Pranav Rajpurkar, and David A Kim. Mc-med, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12(1):1094, 2025.

Min Gu Kim and Jaehyo Jung. A u-net based lsgan framework for 12-lead ecg synthesis in wearable iomt systems. *IEEE Internet of Things Journal*, 2026.

Dominik D Kranz, Jan F Krämer, Oruç Kahriman, Alexander Nelde, Niels Wessel, and Christian Meisel. Exploring latent diffusion models for ecg generation on the minute scale. *Computer Methods and Programs in Biomedicine*, pp. 109138, 2025.

Mahantapas Kundu, Mita Nasipuri, and Dipak Kumar Basu. Knowledge-based ecg interpretation: a critical review. *Pattern Recognition*, 33(3):351–373, 2000.

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

Jun Li, Aaron D Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *Nejm ai*, 2(7):AIoa2401033, 2025a.

Qi Li, Zhenyu Zhang, Lei Yao, Zhaoxia Li, Tianyi Zhong, and Yong Zhang. Diffusion-based decoupled deterministic and uncertain framework for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 67475–67503, 2025b.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. *arXiv preprint arXiv:2403.06659*, 2024.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International conference on learning representations*, volume 2024, pp. 15012–15035, 2024.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Kandukuri Ratna Prakarsha and Gaurav Sharma. Time series signal forecasting using artificial neural networks: An application on ecg signal. *Biomedical Signal Processing and Control*, 76:103705, 2022.

Kristin E Sandau, Marjorie Funk, Andrew Auerbach, Gregory W Barsness, Kay Blum, Maria Cvach, Rachel Lampert, Jeanine L May, George M McDaniel, Marco V Perez, et al. Update to practice standards for electrocardiographic monitoring in hospital settings: a scientific statement from the american heart association. *Circulation*, 136(19):e273–e344, 2017.

Yuqi She, Junhao Huang, Erke Wang, Xianhao Song, Liping Wang, Haotong An, Chen Xu, Han Zhang, and Xuemin Lin. Ecgmind: A foundation model for ecg classification via dynamic energy guidance. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2412–2423, 2026.

Yuxuan Shu, Peter H Charlton, Fahim Kawsar, Jussi Hernesniemi, and Mohammad Malekzadeh. Clef: Clinically-guided contrastive learning for electrocardiogram foundation models. *arXiv preprint arXiv:2512.02180*, 2025.

Zhijiang Tang, Jiaxin Qi, Yuhua Zheng, and Jianqiang Huang. A comprehensive benchmark for electrocardiogram time-series. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6490–6499, 2025.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.

Wei Wang, Jian Chen, Junxin Chen, Zeling Xu, Yuntao Zou, and Henry HY Tong. Sustainable and responsible ecg-based ai diagnostics: Masked frequency reconstruction with peak-aware transformers. In *Proceedings of the ACM Web Conference 2026*, pp. 9200–9210, 2026a.

Xiaoda Wang, Kaiqiao Han, Yuhao Xu, Xiao Luo, Yizhou Sun, Wei Wang, and Carl Yang. Se-diff: Simulator and experience enhanced diffusion model for comprehensive ecg generation. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Yongchao Wang, Fubin Guo, Shizhan Tang, Junchao Fan, Xiuli Bi, and Bin Xiao. Enhancing inter-lead correlations: A novel diffusion gan framework for 12-lead ecg generation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2841–2845. IEEE, 2026c.

Weiwei Ye, Zhuopeng Xu, and Ning Gui. Non-stationary diffusion for probabilistic time series forecasting. *arXiv preprint arXiv:2505.04278*, 2025.

Xiaoyan Yuan, Wei Wang, Han Liu, Jian Chen, and Xiping Hu. Ecg2tok: Ecg pre-training with self-distillation semantic tokenizers. In *IJCAI*, pp. 9990–9998, 2025.

Jianwei Zheng, Hangyuan Guo, and Huimin Chu. A large scale 12-lead electrocardiogram database for arrhythmia study. *PhysioNet*, August 2022. doi: 10.13026/wgex-er52. URL https://doi.org/10.13026/wgex-er52. Version 1.0.0.

A Appendix

A.1 Dataset Curation

ED-ECG-Traj is constructed from continuous Lead II ECG recordings in MC-MED. The curation pipeline converts heterogeneous bedside-monitor recordings into patient-disjoint history–future waveform examples while enforcing temporal continuity, preventing future information leakage, and limiting unequal contributions from patients with long recordings. The complete procedure is summarized in Algorithm 1.

Candidate construction. We first assign patients to training, validation, and test partitions using a 90:5:5 patient-level split, ensuring that no patient appears in more than one partition. Within each continuous recording, candidate forecast origins are identified only when the preceding 300 seconds of ECG and the required future horizon are fully available. We include anchors associated with recorded clinical events together with routine anchors sampled along the monitoring timeline. Each retained origin defines a 300-second historical waveform and future waveforms at the available 10-, 30-, and 60-second horizons.

Waveform preprocessing. All waveforms are sampled at 500 Hz and processed in independent 10-second blocks using a soft zero-phase 0.5–45 Hz band-pass filter with reflected boundary padding. Processing blocks independently prevents samples from a future interval from affecting the filtered representation of the history. A one-sample discrepancy caused by timestamp slicing is deterministically padded or trimmed; larger duration mismatches are treated as incomplete windows and rejected. For each example, the normalization center and scale are estimated exclusively from the filtered 300-second history using the median and robust median absolute deviation. The same parameters are then applied to both the historical and future waveforms. Thus, neither filtering nor normalization introduces information from the prediction interval into the model input.

Technical quality control. Quality control is performed in the same filtered and normalized signal space used for model training and evaluation. We reject windows containing non-finite values, excessive clipping or flatlining, implausible signal variance, or insufficient waveform coverage. Additional artifact indicators measure high-frequency power, spectral entropy, zero-crossing rate, derivative energy, QRS prominence, median-beat consistency, baseline correction magnitude, and residual electrode motion. Data-dependent thresholds are calibrated exclusively on the training partition using robust statistics and are then fixed for validation and test data. A window is rejected when it satisfies a hard failure criterion, accumulates at least four artifact votes, or exhibits severe residual baseline motion. Clinical rhythm outliers are not rejection criteria.

Patient-level sampling and manifests. After quality control, all eligible recorded-event anchors are retained. Routine windows are selected deterministically and approximately uniformly across each patient's timeline to fill a target budget of 10 windows per patient. Patients with more than 10 eligible recorded-event anchors may therefore exceed this target. The selected examples are sorted deterministically and written to split- and horizon-specific manifests. Each manifest records the patient, encounter, waveform segment, forecast origin, available horizons, and preprocessing metadata. All evaluated methods use the same fixed test manifest for each horizon.

Evaluation-only metadata. Clinical weak labels are attached only after waveform curation and are stored in separate sidecar files. They are used for stratified evaluation and are not provided as model inputs. Stable and dynamic strata are likewise computed retrospectively from the observed future waveform and are excluded from sample selection, preprocessing, and training. This separation prevents evaluation annotations from influencing dataset construction or model predictions.

Algorithm 1 Construction of ED-ECG-Traj

Require: Continuous Lead II recordings $\mathcal{R}$; patient identifiers; recorded-event timestamps; history length $L = 300$ s; horizons $\mathcal{H} = \{10, 30, 60\}$ s
Ensure: Patient-disjoint, quality-controlled manifests $\mathcal{M}_{\text{train}}, \mathcal{M}_{\text{val}}, \mathcal{M}_{\text{test}}$
Assign patients to training, validation, and test partitions using a 90:5:5 split
Identify recorded-event and routine candidate origins with at least L seconds of preceding ECG and sufficient future coverage
for each training candidate **do**
    Load the historical waveform and the longest available future waveform
    Canonicalize waveform lengths, allowing only a one-sample pad or trim
    Filter history and future independently in 10-second blocks
    Estimate normalization parameters from the filtered history only
    Normalize both history and future using the historical parameters
    Compute technical quality features
end for
Calibrate data-dependent quality thresholds using training candidates only
for each split $s \in \{\text{train}, \text{val}, \text{test}\}$ **do**
    for each candidate in split s **do**
        Apply the same loading, filtering, and history-only normalization procedure
        Reject incomplete windows, hard signal failures, multi-indicator artifacts, and severe residual motion
        Record all eligible horizons in $\mathcal{H}$
    end for
    for each patient in split s **do**
        Retain all eligible recorded-event anchors
        Fill the remaining target budget with routine windows selected uniformly along the patient timeline
    end for
    Sort examples deterministically and export fixed horizon-specific manifests
end for
Attach clinical weak labels as evaluation-only sidecars
Compute stable and dynamic strata only during retrospective evaluation
Audit patient separation, sample counts, rejection rates, and manifest hashes

A.2 Dynamics Stratification

We partition evaluation windows into stable and dynamic strata using fixed ECG-based criteria. For each trajectory origin, we compare the observed future ECG with a reference continuation constructed by repeating the most recent 10 seconds of historical ECG. All windows are classified using the same criteria. Group assignments are independent of learned-model predictions and are used only for retrospective evaluation, not as model inputs.

Reference construction. Let $x \in \mathbb{R}^T$ denote the observed future ECG and $r \in \mathbb{R}^T$ the Last-10s Repeat reference. The reference is repeated and truncated to match the future duration H. Both waveforms follow the same preprocessing and history-derived normalization protocol. We compare their estimated heart rate, median RR interval, detected peak rate, QRS amplitude, waveform standard deviation, and normalized beat morphology. These descriptors define the strata and are distinct from the set of outcome metrics reported in the main tables.

Dynamic-window criteria. A window is classified as dynamic if at least one criterion in Table 3 is satisfied. Heart rate and median RR interval are computed using the shared evaluation peak detector. Detected peak rate is the number of detected peaks divided by the window duration. QRS amplitude is the median absolute difference between waveform values at detected peaks and the waveform median. Beat morphology distance follows the normalized median-beat definition in Appendix A.3.

Table 3: Fixed criteria for dynamics stratification. A window is classified as dynamic if at least one criterion is satisfied. Here, x denotes the observed future ECG, r the historical repetition reference, and H the future duration in seconds.

Characteristic	Dynamic-window criterion
Heart rate	$\lvert \mathrm{HR}(x) - \mathrm{HR}(r) \rvert \geq 10$ bpm
Median RR interval	$\lvert \mathrm{RR}(x) - \mathrm{RR}(r) \rvert \geq 0.08$ s
Detected peak rate	$\lvert N(x) - N(r) \rvert / H \geq 0.20$ beats/s
QRS amplitude	$\rho(A(x), A(r)) \geq 1.25$
Waveform standard deviation	$\rho(\sigma(x), \sigma(r)) \geq 1.25$
Beat morphology	$\mathrm{Morph}(x, r) \geq 0.25$

For QRS amplitude and waveform standard deviation, the symmetric ratio is defined as:

$$\rho(a, b) = \frac{\max(|a|, |b|)}{\min(|a|, |b|)}. \tag{3}$$

This ratio detects changes in either direction and is evaluated only when both values are finite and $\min(|a|, |b|) > 10^{-12}$. HR, RR-interval, and morphology criteria are evaluated only when the required measurements are finite. An unavailable measurement does not trigger its corresponding criterion.

Stable-window definition. A window is classified as stable if none of the available criteria is satisfied. The two strata are therefore mutually exclusive and collectively cover all evaluated windows. Stable denotes the absence of threshold-crossing changes in the available waveform-derived descriptors; a persistent abnormal rhythm can therefore belong to this stratum. Unavailable physiological measurements may reduce sensitivity to changes.

Evaluation across horizons. We apply the same thresholds across future horizons but determine group membership separately for each horizon. The same trajectory origin may therefore receive different labels for its 10-, 30-, and 60-second future windows. Within each horizon, all methods are evaluated on matched windows using identical group assignments. We report both unique patient counts and window counts for each stratum. Although each window belongs to exactly one stratum, a patient may contribute windows to both, so patient counts are not additive across strata. The original patient-disjoint training, validation, and test splits remain unchanged.

Interpretation. The resulting strata provide an operational characterization of departures from recent ECG persistence under predefined waveform-derived criteria. Because the strata are defined relative to Last-10s Repeat and share HR- and morphology-related descriptors with parts of the evaluation, the stratified results are interpreted as controlled stress tests of persistence-relative waveform modeling.

A.3 Evaluation Metrics Details

We detail the metrics used in our ECG-WorldBench. Let $x \in \mathbb{R}^T$ denote an observed future ECG and $\{\hat{x}^{(k)}\}_{k=1}^K$ denote trajectories generated from the same historical condition. For deterministic methods, $\hat{x}$ denotes the single prediction. For probabilistic methods, pointwise and physiological metrics use the first generated sample, rather than the ensemble-mean waveform or a sample selected using the target. Waveform comparisons are performed on the common 500 Hz evaluation grid. Amplitude-dependent metrics are expressed in the normalized waveform units used by the benchmark.

A.3.1 Waveform accuracy

Mean absolute error (MAE). MAE measures the average absolute deviation between the predicted and observed future waveforms:

$$\mathrm{MAE}(x, \hat{x}) = \frac{1}{T} \sum_{t=1}^{T} |x_t - \hat{x}_t|. \tag{4}$$

Lower values indicate better pointwise agreement.

Root mean squared error (RMSE). RMSE measures the root mean squared waveform deviation and gives greater weight to larger errors:

$$\mathrm{RMSE}(x, \hat{x}) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{x}_t)^2}. \tag{5}$$

Lower values indicate better pointwise agreement.

Waveform correlation (Corr.). We compute the Pearson correlation between the predicted and observed waveforms at their original temporal alignment:

$$\mathrm{Corr}(x, \hat{x}) = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(\hat{x}_t - \bar{\hat{x}})}{\sqrt{\sum_{t=1}^{T}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{T}(\hat{x}_t - \bar{\hat{x}})^2}}. \tag{6}$$

Here, $\bar{x}$ and $\bar{\hat{x}}$ denote temporal means. No additional temporal shift is optimized during evaluation. Higher values indicate stronger waveform correlation. Correlation is undefined when the denominator is numerically negligible.

A.3.2 Physiological Structure

Peak extraction. HR estimation, R-peak F1, and morphology evaluation use the same polarity-invariant peak detector. The waveform is median-centered, and local maxima of its absolute amplitude are selected using a threshold equal to the median absolute amplitude plus 2.5 times its median absolute deviation. A 300 ms refractory interval suppresses nearby candidates, retaining the stronger peak. These automatically extracted peaks serve as QRS proxies rather than manually annotated R peaks. Given that these indicators are subject to the influence of the detector, we have further supplemented the analysis based on the WFDB-XQRS detector, as in Appendix A.4.

HR MAE. For a waveform with detected peak times $\{r_j\}_{j=1}^{J}$, heart rate is estimated from the median inter-peak interval:

$$\mathrm{HR}(x) = \frac{60}{\mathrm{median}_j(r_{j+1} - r_j)}. \tag{7}$$

Peak times are measured in seconds, yielding heart rate in beats per minute. The per-window error is $|\,\mathrm{HR}(x) - \mathrm{HR}(\hat{x})|$, and HR MAE is its aggregate mean. Heart rate is undefined when fewer than two peaks are detected.

R-peak F1. Predicted and reference R peaks are matched one-to-one within a 50 ms tolerance. Let M denote the number of matched pairs, and let N_{pred} and N_{ref} denote the numbers of predicted and reference peaks, respectively. We compute

$$F_1 = \frac{2M}{N_{\mathrm{pred}} + N_{\mathrm{ref}}}. \tag{8}$$

The score is zero when peaks occur on only one side or when no matches are found, and is undefined only when both peak sets are empty; such cases are excluded from the average. Higher values indicate better agreement in R-peak timing.

Beat-aligned morphology distance (Morph.). We extract complete beat segments spanning 200 ms before to 400 ms after each detected peak. For each waveform separately, we compute the pointwise median across these aligned segments to obtain a representative beat template b. We then median-center and RMS-normalize the template:

$$\tilde{b} = \frac{b - \mathrm{median}(b)}{\max\left(\sqrt{\frac{1}{M}\sum_{m=1}^{M}(b_m - \mathrm{median}(b))^2},\ 10^{-8}\right)}. \tag{9}$$

The reported distance is the RMSE between the normalized reference and generated templates:

$$\mathrm{Morph}(x, \hat{x}) = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \left(\tilde{b}_m(x) - \tilde{b}_m(\hat{x}) \right)^2}. \tag{10}$$

Lower values indicate closer normalized beat morphology. This metric does not directly measure absolute beat amplitude or beat-by-beat temporal alignment. It is undefined when either waveform contains no complete extracted beat.

A.3.3 Probabilistic Quality

Energy score. For a predictive distribution F and an observed future waveform y, the energy score is

$$\mathrm{ES}(F, y) = \mathbb{E}\, d(X, y) - \frac{1}{2} \mathbb{E}\, d(X, X'), \tag{11}$$

where X and X' are independent draws from F. For waveforms $a, b \in \mathbb{R}^T$, we use the temporally normalized Euclidean distance

$$d(a, b) = \left(\frac{1}{T} \sum_{t=1}^{T} (a_t - b_t)^2 \right)^{1/2}. \tag{12}$$

Given K generated waveforms $\{\hat{x}^{(k)}\}_{k=1}^{K}$, the reported empirical energy score is

$$\widehat{\mathrm{ES}} = \frac{1}{K} \sum_{k=1}^{K} d(\hat{x}^{(k)}, y) - \frac{1}{2K^2} \sum_{k=1}^{K} \sum_{k'=1}^{K} d(\hat{x}^{(k)}, \hat{x}^{(k')}). \tag{13}$$

Lower values indicate better agreement between the sampled predictive distribution and the observed waveform. For a deterministic prediction, the pairwise dispersion term vanishes and the score reduces to $d(\hat{x}, y)$. For deterministic predictions, the energy score reduces to waveform RMSE and is therefore omitted from the probabilistic comparison to avoid duplicating the pointwise evaluation.

Coverage and interval width. At each future time point, we construct a central 90% prediction interval from the empirical 5th and 95th percentiles of the generated samples:

$$\ell_t = Q_{0.05}\left(\{\hat{x}_t^{(k)}\}_{k=1}^{K} \right), \qquad u_t = Q_{0.95}\left(\{\hat{x}_t^{(k)}\}_{k=1}^{K} \right). \tag{14}$$

Empirical quantiles use linear interpolation. Coverage and interval width are computed as:

$$\widehat{\mathrm{Coverage}}_{5\text{-}95} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{\ell_t \le x_t \le u_t\}, \qquad \mathrm{Width}_{90} = \frac{1}{T} \sum_{t=1}^{T} (u_t - \ell_t). \tag{15}$$

These are pointwise intervals rather than simultaneous prediction bands for complete waveforms. Coverage and interval width are interpreted jointly. With $K = 8$, empirical coverage and width depend on the predictive distribution and quantile estimator; their finite-sample properties are discussed in Appendix A.16.2.

A.3.4 Distributional and Semantic Consistency

Feature Fréchet distance (FFD). We use the frozen pretrained FFD feature extractor supplied by ECG-Benchmark. Waveforms are resampled to its 100 Hz input rate, and output tokens are averaged to obtain one feature vector per trajectory. Let μ_x, Σ_x and $\mu_{\hat{x}}, \Sigma_{\hat{x}}$ denote the empirical feature means and covariance matrices of the observed and generated waveform collections. For feature dimension d, the reported metric is:

$$\mathrm{FFD} = \frac{1}{\sqrt{d}} \left[\|\mu_x - \mu_{\hat{x}}\|_2^2 + \mathrm{Tr}\Big(\Sigma_x + \Sigma_{\hat{x}} - 2(\Sigma_x^{1/2} \Sigma_{\hat{x}} \Sigma_x^{1/2})^{1/2} \Big) \right]. \tag{16}$$

The factor $1/\sqrt{d}$ follows the benchmark implementation. Lower values indicate closer feature distributions. Unlike paired waveform metrics, FFD measures collection-level distributional agreement.

ECGFounder and CLEF cosine similarities. We separately extract representations using frozen ECGFounder and CLEF encoders. For each encoder f, semantic similarity is measured by:

$$\mathrm{Cos}_f(x, \hat{x}) = \frac{f(x)^\top f(\hat{x})}{\|f(x)\|_2 \|f(\hat{x})\|_2}. \tag{17}$$

Trajectories are divided into non-overlapping 10-second blocks, and cosine similarities are computed for corresponding observed and generated blocks and averaged within each trajectory. For probabilistic methods, these reported similarities use the first generated sample. Higher values indicate greater representation-level agreement between generated and observed ECGs under the corresponding frozen encoder.

ECGFounder probability MAE. In addition to embedding similarity, we compare ECGFounder classification probabilities over the predefined primary single-lead label subset $\mathcal{C}$. Let $p_c(x)$ denote the sigmoid probability for label c. For each corresponding 10-second block pair, we compute:

$$\mathrm{ProbMAE}(x, \hat{x}) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} |p_c(x) - p_c(\hat{x})|. \tag{18}$$

Block-level values are averaged within each trajectory. The subset is fixed by the evaluator's primary single-lead label mapping, rather than selected from test performance. Lower values indicate closer ECGFounder output probabilities between generated and observed ECGs, measuring model-output consistency.

Aggregation and undefined values. Unless otherwise specified, per-window metrics are averaged over finite values across evaluated windows. Coverage and interval width are first averaged within each patient and then across patients with equal patient weights. FFD is computed from the pooled feature collections. Patient-cluster bootstrap inference resamples patients while retaining their associated windows and is distinct from patient-equal aggregation. Undefined physiological measurements are not replaced with zero; their availability should be considered when interpreting comparisons.

A.4 Sensitivity to QRS Detector Choice

The primary evaluation uses a lightweight QRS detector with a 300 ms refractory interval. Since heart-rate MAE, R-peak F1, beat-aligned morphology distance, and parts of the dynamics stratification depend on detected peaks, we conduct an additional evaluation using WFDB-XQRS to examine sensitivity to detector choice. XQRS uses a 200 ms refractory interval and a maximum-heart-rate configuration of 300 bpm. No models are retrained, and both detectors evaluate the same fixed primary prediction for each of the 2,954 test windows at every horizon. The original detector remains the main evaluation protocol.

We recompute the three physiological metrics while retaining the original median-RR heart-rate definition, 50 ms peak-matching tolerance, and morphology normalization. Stable/dynamic labels are also recomputed using the same thresholds and saved Last-10s Repeat reference, replacing only detector-dependent features. Table 4 reports the XQRS measurements, and Table 5 reports the proportion of windows for which each metric can be computed.

Physiological measurements. Under XQRS, PSLD achieves heart-rate MAEs of 5.1551, 4.2200, and 3.5630 bpm at 10, 30, and 60 s, respectively, with morphology distances of 0.2369, 0.2476, and 0.2184. Last-10s Repeat retains lower heart-rate and morphology errors, whereas PSLD achieves higher R-peak F1 at every horizon. PSLD obtains the highest F1 at 10 s, but the ordering changes at longer horizons. At 60 s, its F1 of 0.1802 is below those of Phase-Aligned Median Beat, PatchTST, Latent-NsDiff, and Latent-D3U, indicating greater detector sensitivity in long-horizon peak alignment.

Metric computability. Heart-rate estimation requires at least two detected peaks, while morphology comparison requires complete beat segments. Windows without the required detections receive undefined measurements and are excluded from the corresponding means, rather than assigned zero.

Table 4: Physiological evaluation using WFDB-XQRS. Means are computed over each method's computable windows; corresponding proportions are reported in Table 5.

Horizon	Metric	Last-10s Repeat	Phase-Aligned Median Beat	PatchTST	PSSM	Latent-NsDiff	Latent-D3U	PSLD
10 s	HR MAE ↓	4.3371	30.4839	94.5008	82.1881	100.2903	99.7903	5.1551
	R-peak F1 ↑	0.1783	0.3417	0.2098	0.1617	0.2125	0.2115	0.3994
	Morph. ↓	0.1724	0.4735	1.3883	1.3627	1.3840	1.3800	0.2369
30 s	HR MAE ↓	3.4636	30.3355	97.0454	96.9572	98.7408	98.7686	4.2200
	R-peak F1 ↑	0.1643	0.2457	0.2012	0.0956	0.2026	0.2027	0.2214
	Morph. ↓	0.1412	0.4568	1.4052	1.3986	1.3644	1.3719	0.2476
60 s	HR MAE ↓	3.3786	30.2574	98.0260	107.1497	99.5345	99.0460	3.5630
	R-peak F1 ↑	0.1602	0.2072	0.2001	0.0321	0.2004	0.1996	0.1802
	Morph. ↓	0.1349	0.4524	1.4129	1.3645	1.3490	1.3613	0.2184

Table 5: Proportion of test windows with computable XQRS-based metrics (%). Percentages indicate the fraction of test windows yielding a finite value for each metric.

Horizon	Metric	Last-10s Repeat	Phase-Aligned Median Beat	PatchTST	PSSM	Latent-NsDiff	Latent-D3U	PSLD
10 s	HR MAE	100.0	100.0	100.0	85.1	100.0	100.0	100.0
	R-peak F1	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Morph.	100.0	100.0	100.0	85.5	100.0	100.0	100.0
30 s	HR MAE	100.0	100.0	100.0	55.5	100.0	100.0	100.0
	R-peak F1	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Morph.	100.0	100.0	100.0	55.9	100.0	100.0	100.0
60 s	HR MAE	100.0	100.0	100.0	20.5	100.0	100.0	100.0
	R-peak F1	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Morph.	100.0	100.0	100.0	20.5	100.0	100.0	100.0

R-peak F1 remains defined as zero when peaks are detected in only one of the two waveforms, explaining why its computability can exceed that of the other metrics. PSSM's XQRS-based heart-rate computability decreases from 85.1% at 10 s to 55.5% at 30 s and 20.5% at 60 s, while PSLD retains 100% computability for all three metrics. These proportions quantify the availability of each metric under XQRS and specify the evaluation coverage of the means reported in Table 4.

Ranking and patient-level comparisons. To compare rankings without differences in metric availability, we use windows with computable measurements for all methods under both detectors. Spearman rank correlations for heart-rate MAE, R-peak F1, and morphology distance are 0.857, 0.679, and 0.750 at 10 s; 0.679, 0.536, and 0.964 at 30 s; and 0.607, 0.286, and 0.893 at 60 s. The 60-second heart-rate and morphology comparisons include 606 jointly computable windows, so these rankings describe that subset rather than the full cohort. For patient-level comparisons, we average paired window differences within each patient and then weight patients equally. On identical jointly computable windows, all 27 overall mean-difference directions are preserved when comparing PSLD with PSSM, Phase-Aligned Median Beat, and Last-10s Repeat across three metrics and three horizons.

Dynamics stratification. Replacing the detector changes the number of dynamic windows from 949 to 693 at 10 s, from 750 to 538 at 30 s, and from 736 to 501 at 60 s. Stratified paired comparisons therefore use windows assigned to the same stratum by both detectors. We estimate 95% confidence intervals using 2,000 patient-bootstrap replicates and apply two-sided Wilcoxon signed-rank tests with Holm correction across the reported comparisons within each horizon. Three dynamic-stratum F1 comparisons change mean-difference direction: PSLD versus Last-10s Repeat at 30 and 60 s, and PSLD versus PSSM at 60 s. The former two are not significant after correction under either

detector; the latter changes from a small, non-significant disadvantage to a significant advantage under XQRS.

Summary. The directions of the selected overall patient-level comparisons are preserved across detectors, while long-horizon R-peak F1 rankings, metric computability, and dynamics stratification vary. These findings support cross-detector consistency of the selected overall comparisons, while highlighting the sensitivity of peak-alignment and stratified results to the detection protocol. "'

A.5 More Details About Proposed Framework.

We introduce a two-stage model for ECG history-conditioned trajectory generation. A physiology-supervised variational autoencoder (VAE) first learns a compact representation of ECG waveforms. A conditional latent diffusion model then generates future trajectories using multiscale historical context. Physiological supervision shapes the representation during VAE pre-training, while stochastic latent generation allows variation beyond periodic continuation.

Physiology-supervised latent representation. The VAE encoder parameterizes a diagonal Gaussian posterior,

$$q_\phi(z \mid x) = \mathcal{N}\big(\mu_\phi(x), \mathrm{diag}(\sigma_\phi^2(x))\big), \qquad \hat{x} = D_\psi(z), \tag{19}$$

where each latent token has 48 dimensions. A physiology prediction head maps the full posterior mean to waveform-derived physiological descriptors, including cardiac phase, rhythm, amplitude, and beat morphology. This supervision encourages physiologically informative representations across the full latent space. We train the VAE using

$$\mathcal{L}_{\mathrm{VAE}} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}} + \beta\mathcal{L}_{\mathrm{KL}}, \tag{20}$$

where $\mathcal{L}_{\mathrm{rec}}$ is waveform ℓ_1 reconstruction loss, $\mathcal{L}_{\mathrm{phys}}$ is a masked Smooth-ℓ_1 loss on normalized physiological descriptors, and $\mathcal{L}_{\mathrm{KL}}$ regularizes the posterior toward a standard Gaussian using free-nats regularization and KL warm-up. After pre-training, the encoder and decoder remain frozen. We use posterior means as latent representations and apply a normalization transform $\mathcal{W}$ whose statistics are estimated exclusively from the training split.

Multiscale historical conditioning. Given an observed ECG history x^{hist}, we obtain its normalized latent representation,

$$z^{\mathrm{hist}} = \mathcal{W}\big(\mu_\phi(x^{\mathrm{hist}})\big), \tag{21}$$

and construct historical conditioning as

$$c_{\mathrm{hist}} = \big\{C_{\mathrm{multi}}(z^{\mathrm{hist}}), P_{\mathrm{recent}}(z^{\mathrm{hist}}), r_{\mathrm{hist}}\big\}. \tag{22}$$

Here, C_{multi} summarizes up to 300 seconds of history using finer temporal resolution near the prediction boundary and coarser resolution for more distant context. P_{recent} retains the most recent 10 seconds as clean prefix tokens, while r_{hist} provides history-derived rhythm and boundary-phase information. These features condition a temporal diffusion transformer through historical memory and prefix tokens. All conditioning is derived exclusively from observed ECG history; future ECGs and evaluation-stratum labels are not used as conditioning inputs.

Conditional latent diffusion. During training, the observed future ECG is encoded into a normalized clean latent trajectory,

$$z_0 = \mathcal{W}\big(\mu_\phi(x^{\mathrm{future}})\big), \tag{23}$$

which is perturbed through a forward diffusion process:

$$z_\tau = \alpha_\tau z_0 + \sigma_\tau \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{24}$$

where τ denotes diffusion time and $\alpha_\tau^2 + \sigma_\tau^2 = 1$. The temporal diffusion transformer predicts the velocity target

$$v_\tau = \alpha_\tau \epsilon - \sigma_\tau z_0, \tag{25}$$

using the objective

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{z_0, c_{\mathrm{hist}}, \tau, \epsilon}\left[\big\|v_\theta(z_\tau, \tau, c_{\mathrm{hist}}) - v_\tau\big\|_2^2\right]. \tag{26}$$

The corresponding clean-latent estimate is $\hat{z}_0 = \alpha_\tau z_\tau - \sigma_\tau v_\theta$. We additionally use latent self-conditioning, supplying a detached clean-latent estimate as an auxiliary denoising input. Together, physiology-supervised representation learning and history-conditioned diffusion support stochastic generation in a physiologically informed latent space.

Stochastic trajectory generation. At inference, independently sampled Gaussian noise is iteratively denoised using DDIM sampling under the same historical conditioning. After reversing latent normalization, the frozen VAE decoder maps the generated latents into future ECG waveforms:

$$\hat{x}^{\text{future},(k)} = D_\psi\left(\mathcal{W}^{-1}(\hat{z}_0^{(k)})\right), \qquad k = 1, \ldots, K. \tag{27}$$

These samples represent alternative continuations of the same observed history, allowing variation in future rhythm and morphology rather than prescribing a repeated historical waveform.

Implementation details. The Stage-1 VAE uses three convolutional encoder and decoder stages with channel widths 64, 128, and 256, producing 48-dimensional latent tokens at 10 Hz. Its physiological head predicts 26 waveform-derived descriptors. We set $\lambda_{\text{phys}} = 0.05$, $\beta = 10^{-3}$, and the free-nats threshold to 0.02 per latent dimension, with KL warm-up over 10 epochs. The VAE is trained for 50 epochs using AdamW with a learning rate of 2×10^{-4}, weight decay of 10^{-4}, and a per-GPU batch size of 32. The Stage-2 denoiser is a 12-block temporal DiT with width 768 and 12 attention heads, conditioned on 300 seconds of multiscale history and a 10-second clean prefix. It uses 1,000 diffusion steps, velocity prediction, a self-conditioning probability of 0.5, and history dropout of 0.05. Stage 2 is trained for 20 epochs using AdamW with the same learning rate and weight decay, an effective batch size of 64.

A.6 Baseline Details

We evaluate six baselines spanning training-free periodic continuation, deterministic waveform forecasting, and probabilistic latent-space generation. All methods use the benchmark's patient-disjoint splits and are evaluated on matched windows against the same observed future ECG targets. Learned methods receive 300 seconds of historical ECG, with separate models trained for the 10-, 30-, and 60-second forecasting horizons. The periodic baselines retain their recent-waveform or recent-beat construction rules rather than mechanically repeating the full historical context. PatchTST and PSSM operate in waveform space, whereas Latent-NsDiff and Latent-D3U use the same frozen Stage 1 representation as our model.

A.6.1 Training-Free Periodic Baselines

Last-10s Repeat. This baseline uses the most recent 10 seconds of preprocessed historical ECG as a fixed waveform template. It copies this segment for the 10-second horizon and repeats and truncates it for longer horizons. The template remains unchanged throughout generation, and no future observations are used. This provides a direct reference for continuation of the patient's recent waveform without learned trajectory modeling.

Phase-Aligned Median Beat. This baseline estimates a patient-specific beat template from the most recent 30 seconds of historical ECG. Following peak detection, inter-peak intervals between 0.25 and 2.5 seconds are retained, and their median determines the continuation period. Valid complete peak-to-peak cycles are resampled to this period and combined using a pointwise median. The resulting template is repeated with its initial phase determined by the elapsed time since the last historical peak. At least three valid cycles are required; otherwise, the method falls back to repeating a recent historical waveform segment. Both template estimation and phase alignment depend exclusively on observed history.

A.6.2 Deterministic Learned Baselines

PatchTST Nie et al. (2022). We adapt the official supervised PatchTST implementation to single-lead ECG forecasting. The model partitions the resulting sequence into overlapping patches and uses a Transformer backbone to predict the requested future waveform. Predictions are restored to 500 Hz for evaluation. The configuration uses six Transformer layers; four attention heads; a hidden dimension of 64; a feed-forward dimension of 256; and backbone dropout of 0.1. Reversible instance normalization is enabled. Training minimizes waveform mean squared error at the model sampling rate. PatchTST does not access the Stage 1 encoder, decoder, or latent cache.

PSSM Tang et al. (2025). We use the official PSSM backbone from ECG-Benchmark with a forecasting head. The backbone processes the historical single-lead waveform, and a two-layer multilayer perceptron maps its temporal output to the requested future length. The backbone model dimension is 32, and the forecasting head has a hidden dimension of 1024, a ReLU activation, and dropout of 0.2. Training minimizes waveform mean squared error at 100 Hz, without using the Stage 1 representation.

A.6.3 Probabilistic Latent-Space Baselines

Shared latent-space adaptation. Directly modeling long ECG histories and generating extended high-resolution waveforms with the original diffusion architectures would incur substantial computational and memory costs. To support multi-horizon waveform generation conditioned on long histories, we adapt NsDiff and D3U to the shared frozen Stage 1 VAE latent space. The representation has 48 channels and a temporal resolution of 10 latent tokens per second. Consequently, the 300-second history contains 3,000 latent tokens, while the 10-, 30-, and 60-second targets contain 100, 300, and 600 tokens, respectively. Latents are standardized using fixed statistics estimated exclusively from the training split and shared across latent-space methods for each horizon. Generated latents are transformed back to the decoder's original scale and decoded into 500 Hz ECG waveforms. The VAE remains frozen throughout baseline training. We refer to these adaptations as Latent-NsDiff and Latent-D3U to distinguish them from the original implementations.

Latent-NsDiff Ye et al. (2025). We retain the official conditional mean predictor, history-dependent variance estimator, and diffusion network, adapting their input and output channels to the 48-dimensional latent representation. Historical latents provide the conditioning sequence, while the decoder input concatenates the most recent half of the historical sequence with zero placeholders for future positions. Temporal covariates are set to zero. The backbone uses a hidden dimension of 128, four attention heads, three encoder layers, two decoder layers, a feed-forward dimension of 256, and dropout of 0.05. The variance estimator uses a hidden dimension of 256 and a rolling window of 50 latent tokens. Training combines four equally weighted components: conditional-mean MSE, scale MSE, diffusion noise-prediction MSE, and a variance-ratio objective. The scale term compares the square roots of the predicted and target variances. For the variance ratio $r = \widetilde{\sigma}^2/\widehat{\sigma}^2$, the variance objective is $\mathrm{mean}(r - \log r)$. Diffusion uses 20 steps with a linear noise schedule from $\beta = 10^{-4}$ to 10^{-2}, and inference follows the full reverse chain.

Latent-D3U Li et al. (2025b). We adapt the official conditional forecasting and diffusion modules to multivariate Stage 1 latents while retaining channel-independent patch processing. A patch-based conditional model predicts the future latent mean, and the diffusion model predicts noise added to the residual between this mean and the target latents. The conditional model uses four layers, four attention heads, a hidden dimension of 64, a feed-forward dimension of 256, dropout of 0.05, and a patch length and stride of 30 and 15 latent tokens, respectively. The PatchDN denoiser uses four layers, four attention heads, a hidden dimension of 64, and a patch length and stride of 10 and 5 tokens. Conditional-mean MSE and residual noise-prediction MSE are summed with equal weights. The diffusion process uses 20 steps with a linear noise schedule from $\beta = 10^{-4}$ to 2×10^{-2}. Inference uses the implementation's accelerated sampler with 10 sampling steps and $\eta = 1$, and the sampled residuals are added to the conditional mean before waveform decoding.

Scope of the shared latent-space comparison. Using the same frozen Stage 1 VAE controls latent dimensionality, temporal compression, normalization, and waveform decoding across PSLD, Latent-NsDiff, and Latent-D3U, enabling a controlled comparison of their generative architectures within a common representation. Latent-NsDiff and Latent-D3U are adaptations evaluated in the same physiology-supervised latent space as PSLD. This comparison fixes the representation and decoding pipeline while retaining each method's generative architecture and training objective.

A.6.4 Optimization, Model Selection, and Reproducibility

Optimization and training budget. All learned baselines use AdamW with a constant learning rate of 10^{-4}, weight decay of 10^{-4}, default momentum coefficients $(0.9, 0.999)$, and gradient-norm clipping at 1.0. Training is limited to 20 epochs, with one optimizer update per minibatch and no gradient accumulation. Thus, the budget is specified in epochs rather than as an identical number of

updates across methods. For N_{train} training windows, G distributed workers, and per-worker batch size b, the number of updates per epoch is $\lceil \lceil N_{\text{train}}/G \rceil / b \rceil$ under the distributed sampler used here. The released configurations specify per-GPU training batch sizes of 16, 8, 8, and 8 for PatchTST, PSSM, Latent-NsDiff, and Latent-D3U, respectively; explicit runtime overrides are recorded in the resolved run configurations. Bfloat16 mixed precision is enabled in these configurations. The methods share a maximum budget of 20 epochs, with method-specific batch sizes determining the number of optimizer updates; inference costs are reported in Appendix A.7.

Checkpoint selection. Validation is performed after every epoch. All checkpoints are selected using validation waveform MSE. A checkpoint is updated when the validation objective improves by more than 10^{-6}, and training stops after five consecutive epochs without such an improvement. Final test evaluation uses the selected checkpoint rather than the last training epoch. Test metrics are not used for checkpoint selection.

Randomness and sampling. The base random seed is 42. Distributed workers initialize their random-number generators using the base seed plus their rank, and the training sampler uses the base seed. Each probabilistic baseline generates eight trajectories per test condition. Evaluation seeds are derived deterministically from the sample identifier, draw index, and base seed of 42, making sample-level random draws independent of batching and worker assignment. The eight trajectories are sampled from the same selected checkpoint to characterize within-condition predictive variability.

Evaluation and implementation provenance. All predictions are evaluated at 500 Hz using the common waveform, physiological, probabilistic, distributional, and semantic protocol described in Appendix A.3. The latent baselines retain their own conditioning and diffusion mechanisms rather than using our multiscale history encoder. To make the comparison reproducible, the accompanying code provides the model configurations, pinned upstream revisions, preprocessing and rate-conversion implementations, and resolved run configurations.

A.6.5 Periodic Baselines as Distributional and Semantic References

Role as persistence references. We report FFD, ECGFounder, and CLEF results for the periodic baselines separately from the learned conditional generators. Because these methods explicitly reuse recent waveforms or beat templates, their feature-space performance quantifies the degree to which patient-specific morphology and encoder representations persist into the observed future. The periodic results therefore provide representation-level persistence references that complement the paired waveform, physiological, and stratified evaluations.

Results across forecasting horizons. Table 6 reports results on 2,954 test windows for each forecasting horizon. Last-10s Repeat achieves lower FFD and probability MAE, and higher ECGFounder and CLEF similarities, than Phase-Aligned Median Beat at all three horizons. Its ECGFounder similarity decreases from 0.9236 at 10 seconds to 0.9142 at 60 seconds, while CLEF similarity decreases from 0.9543 to 0.9496. Over the same horizons, FFD increases from 0.0509 to 0.1890 and probability MAE increases from 0.0215 to 0.0230. These results are consistent with substantial persistence of encoder-level ECG characteristics over the evaluated intervals, although distributional and semantic agreement gradually deteriorates with increasing forecast duration.

Implications for benchmark interpretation. At the 10-second horizon, Last-10s Repeat achieves ECGFounder and CLEF similarities of 0.9236 and 0.9543, respectively, while its waveform correlation is 0.0078. This combination separates representation-level persistence from temporal correspondence with the realized future waveform. FFD further characterizes collection-level feature distributions, whereas paired waveform and physiological metrics characterize condition-specific agreement. Together, these dimensions distinguish the ECG properties retained through historical persistence from those captured through conditional future waveform modeling.

Table 6: Distributional and semantic consistency of training-free periodic baselines across forecasting horizons, evaluated on 2,954 test windows per horizon. Results are reported separately as persistence references and are not included in a pooled ranking with learned conditional generators. ECGFounder and CLEF denote embedding cosine similarity; Prob. MAE measures ECGFounder probability disagreement over the primary single-lead label group.

Method	FFD ↓	ECGFounder ↑	CLEF ↑	Prob. MAE ↓
10-second future horizon				
Last-10s Repeat	0.0509	0.9236	0.9543	0.0215
Phase-Aligned Median Beat	0.7553	0.6992	0.8244	0.0749
30-second future horizon				
Last-10s Repeat	0.1237	0.9185	0.9519	0.0224
Phase-Aligned Median Beat	0.9265	0.6977	0.8224	0.0752
60-second future horizon				
Last-10s Repeat	0.1890	0.9142	0.9496	0.0230
Phase-Aligned Median Beat	0.9867	0.6957	0.8209	0.0757

Table 7: Inference latency and peak GPU memory allocation with 300-second historical input and a per-GPU batch size of one. Latency is reported in milliseconds per condition, including all K generated waveforms. Periodic baselines are omitted.

		10s horizon		30s horizon		60s horizon	
Method	K	Time (ms)	Memory (MiB)	Time (ms)	Memory (MiB)	Time (ms)	Memory (MiB)
PatchTST	1	9.90	158.6	10.47	451.2	10.88	890.8
PSSM	1	9.99	206.5	10.22	218.1	8.81	236.4
Latent-NsDiff	1	75.09	452.7	75.37	453.0	76.63	453.4
Latent-D3U	1	89.27	126.1	89.92	140.3	89.44	168.4
PSLD (Ours)	1	560.26	753.3	564.58	753.3	684.75	763.5
Latent-NsDiff	8	559.50	455.6	562.45	456.2	567.35	457.2
Latent-D3U	8	658.77	129.4	661.28	150.8	662.88	193.9
PSLD (Ours)	8	4444.27	753.3	4433.66	753.3	5399.58	764.3

A.7 Inference Time Comparison

We compare the inference latency of learned models at 10-, 30-, and 60-second future horizons. For each setting, we select 32 test history conditions and distribute them across four NVIDIA GeForce RTX 4090 GPUs, with each GPU running an independent model instance at a batch size of one. After five warm-up runs per GPU, each condition is timed five times, and measurements are pooled to obtain the mean latency per condition. Deterministic models produce one trajectory per invocation, while probabilistic models are evaluated separately with one and eight trajectories per condition. Timing starts from preprocessed CPU history waveforms and includes required data transfers, online history encoding, conditioning feature construction, trajectory generation, waveform decoding, and output transfer back to the CPU. Disk I/O, filtering and normalization, model loading, warm-up, and metric computation are excluded. CUDA synchronization is performed at timing boundaries to ensure GPU operations have completed. We also report peak GPU memory allocated during the benchmark. Multiple GPUs are used only to execute independent measurements in parallel, not to accelerate an individual inference request.

Table 7 shows that deterministic models have the lowest latency, requiring approximately 9–11 ms per condition. For single-trajectory generation, Latent-NsDiff and Latent-D3U require approximately 75–77 ms and 89–90 ms, respectively, whereas PSLD requires 560–685 ms. Generating eight trajectories increases PSLD's latency to 4.43–5.40 seconds, with peak GPU memory allocation remaining below 765 MiB. Under the current implementation, trajectories are sampled sequentially, so increasing the sampling budget primarily increases latency rather than peak memory. PSLD's la-

Table 8: Normalized rank scores across evaluation metrics and future horizons. Higher is better.

Method	10 s	30 s	60 s	Overall
PatchTST	0.375	0.475	0.488	0.446
PSSM	0.425	0.425	0.413	0.421
Latent-NsDiff	0.375	0.438	0.475	0.429
Latent-D3U	0.475	0.338	0.325	0.379
PSLD	**0.850**	**0.825**	**0.800**	**0.825**

tency remains below the corresponding future duration at both sampling budgets, although it incurs a higher computational cost than the evaluated baselines.

A.8 Normalized Rank-Based Aggregate Comparison

To summarize performance across evaluation dimensions, we construct a normalized rank-based score for the five learned methods: PatchTST, PSSM, Latent-NsDiff, Latent-D3U, and PSLD. We include the ten metrics available for all five methods: MAE, RMSE, waveform correlation, heart-rate MAE, R-peak F1, morphology distance, FFD, ECGFounder cosine similarity, ECGFounder probability MAE, and CLEF-Small cosine similarity. Energy score, coverage, and interval width are excluded from the common metric set. Coverage and width require multiple samples, whereas the degenerate-distribution energy score for deterministic methods is omitted because it provides no additional distributional information.

For each metric at each future horizon, methods are ranked according to the preferred direction of the metric. The rank is converted into a normalized score:

$$s_{m,h,j} = \frac{M - r_{m,h,j}}{M - 1}, \tag{28}$$

where M is the number of learned methods and $r_{m,h,j}$ is the rank of method m for metric j at horizon h. The highest-ranked method receives a score of 1 and the lowest-ranked method receives 0, independent of the number of evaluated methods. Ties are assigned their average rank. The aggregate score is then computed as

$$S_m = \frac{1}{|\mathcal{H}||\mathcal{J}|} \sum_{h \in \mathcal{H}} \sum_{j \in \mathcal{J}} s_{m,h,j}, \tag{29}$$

where $\mathcal{H} = \{10, 30, 60\}$ denotes the future horizons and $\mathcal{J}$ denotes the ten shared metrics.

As shown in Table 8, PSLD achieves the highest normalized score at every future horizon and ranks first overall. Its score decreases moderately as the future horizon increases but remains substantially higher than those of the other learned methods. Among the remaining methods, the aggregate scores are relatively close, reflecting their different strengths across pointwise, physiological, distributional, and semantic metrics.

Normalization prevents the score range from increasing mechanically with the number of baselines. However, the scores remain relative to the evaluated method set, and rank aggregation discards differences in metric magnitude. The aggregate score summarizes relative rankings across the evaluated methods and complements the metric-specific results.

A.9 Additional Results across Physiological Dynamics

We report the complete six-metric evaluation across the stable and dynamic strata defined in Section 3.2. The strata are determined retrospectively from observed ECG changes and are independent of recorded actions and model predictions. All methods share the same evaluation windows within each horizon, although metric-specific valid-window counts may differ. Stratum membership is determined separately at each horizon, so the stable and dynamic subsets can differ across horizons.

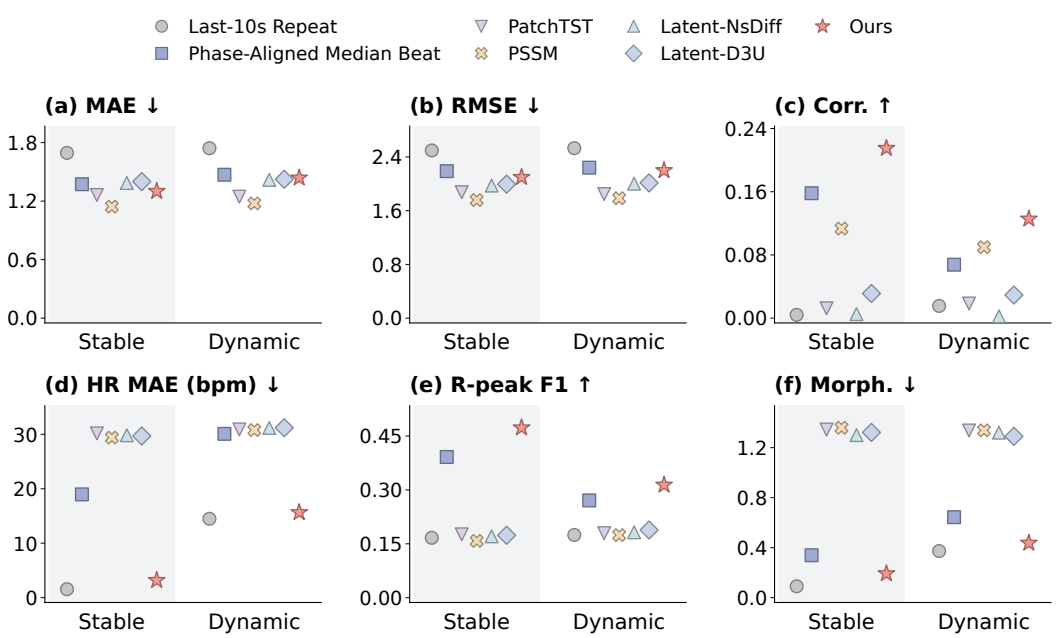


Figure 5: Six-metric evaluation across stable and dynamic ECG windows at the 10-second horizon. The stable stratum contains 2,005 windows from 379 patients, and the dynamic stratum contains 949 windows from 288 patients. Shading denotes the stable stratum. Points represent means over valid windows, and arrows indicate the preferred direction. Patients may contribute to both strata, and metric-specific valid-window counts may differ across methods.

A.9.1 10-Second Future Horizon

Figure 5 shows a clear distinction between pointwise accuracy and physiological fidelity. PSSM achieves the lowest MAE and RMSE in both strata. PSLD, by contrast, achieves the highest waveform correlation and R-peak F1, together with the lowest HR MAE and morphology distance among learned methods. Its correlation decreases from 0.215 in stable windows to 0.126 in dynamic windows, while R-peak F1 decreases from 0.473 to 0.314.

The physiological metrics show a similar degradation under dynamic conditions. For PSLD, HR MAE increases from 3.20 to 15.65 bpm, and morphology distance increases from 0.193 to 0.437. Last-10s Repeat retains the lowest overall HR MAE and morphology distance in both strata, illustrating the strength of recent-pattern continuation for preserving local rhythm and morphology despite its limited agreement with the observed future waveform.

A.9.2 30-Second Future Horizon

The trade-off between pointwise accuracy and physiological fidelity persists at 30 seconds, as shown in Figure 6. PSSM again achieves the lowest MAE and RMSE in both strata. Among learned methods, PSLD achieves the highest correlation and R-peak F1 and the lowest HR MAE and morphology distance. Its correlation decreases from 0.058 in stable windows to 0.036 in dynamic windows, while R-peak F1 decreases from 0.242 to 0.198.

Comparisons with the periodic baselines remain metric-dependent. Phase-Aligned Median Beat achieves higher correlation and R-peak F1 than PSLD in stable windows and a higher R-peak F1 in dynamic windows, whereas PSLD achieves the highest correlation in the dynamic stratum. Last-10s Repeat retains the lowest morphology distance in both strata and the lowest HR MAE in stable windows. In dynamic windows, PSLD and Last-10s Repeat obtain similar HR MAE values of 15.66 and 16.00 bpm, respectively.

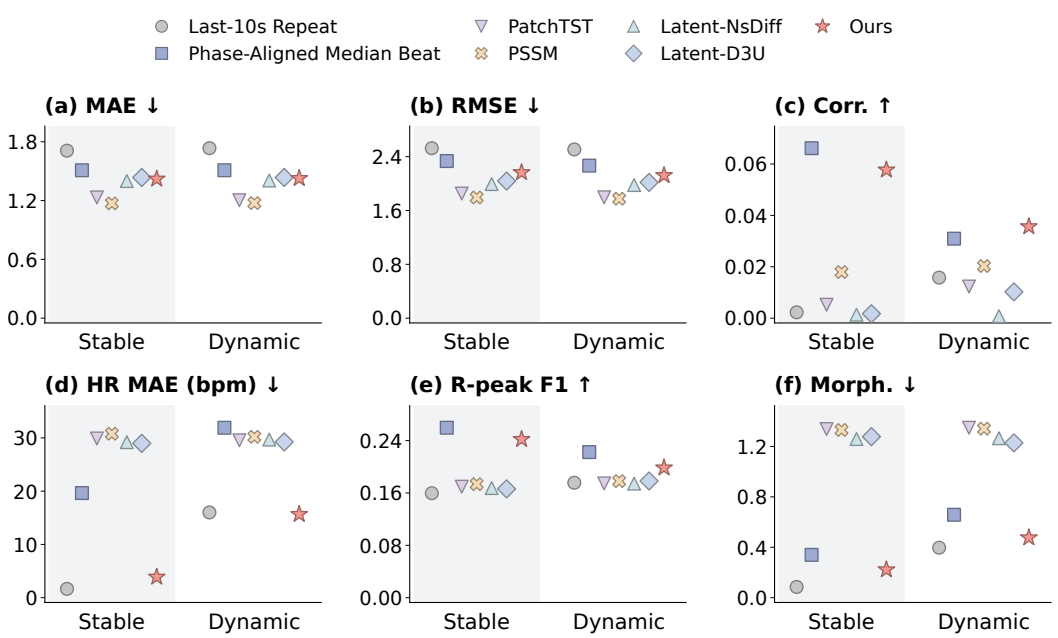


Figure 6: Six-metric evaluation across stable and dynamic ECG windows at the 30-second horizon. The stable stratum contains 2,204 windows from 389 patients, and the dynamic stratum contains 750 windows from 247 patients. Shading denotes the stable stratum. Points represent means over valid windows, and arrows indicate the preferred direction. Patients may contribute to both strata, and metric-specific valid-window counts may differ across methods.

A.9.3 60-Second Future Horizon

Figure 7 shows that the same distinction remains at the longest horizon. PSSM achieves the lowest MAE and RMSE in both strata. Among learned methods, PSLD achieves the highest waveform correlation and R-peak F1 and the lowest HR MAE and morphology distance. Its correlation decreases from 0.025 in stable windows to 0.017 in dynamic windows, while R-peak F1 decreases from 0.194 to 0.178. HR MAE increases from 3.24 to 14.28 bpm, and morphology distance increases from 0.197 to 0.425.

The periodic baselines remain competitive. Phase-Aligned Median Beat achieves higher R-peak F1 than PSLD in both strata and higher correlation in stable windows. PSLD has the highest correlation in dynamic windows, although the absolute difference from the periodic baselines is small. Last-10s Repeat achieves the lowest HR MAE and morphology distance in stable windows and retains the lowest morphology distance in dynamic windows, whereas PSLD achieves the lowest dynamic-window HR MAE. The low absolute correlations across all methods indicate that precise temporal agreement with the observed ECG becomes increasingly difficult over the 60-second horizon.

Across all three horizons, PSLD maintains the strongest physiological fidelity among learned methods, but its performance consistently deteriorates from stable to dynamic windows. Periodic continuation remains particularly competitive for recent rhythm and morphology preservation, while deterministic prediction retains its advantage in pointwise errors. These findings reinforce that performance under physiological dynamics depends strongly on the evaluation dimension and cannot be summarized by a single waveform metric.

A.10 Performance across Clinical Weak-Label Subgroups

We further examine modeling performance within five clinical weak-label subgroups: acute stroke, acute coronary syndrome (ACS), arrhythmia, heart failure, and sepsis. The positive subgroups contain 31 windows from 7 patients for stroke, 26 from 9 for ACS, 63 from 13 for arrhythmia, 58

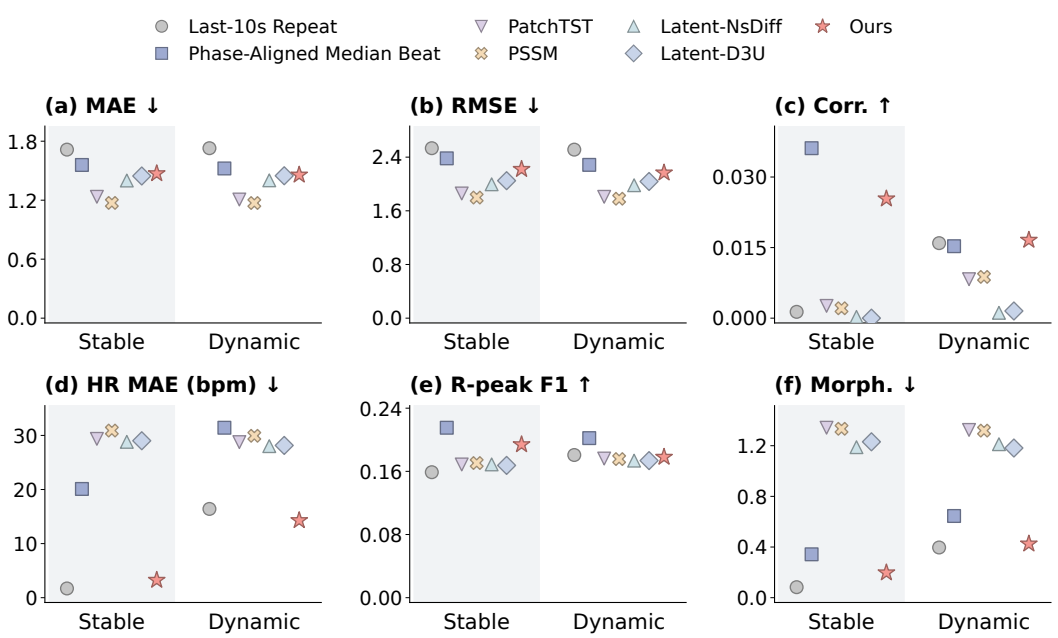


Figure 7: Six-metric evaluation across stable and dynamic ECG windows at the 60-second horizon. The stable stratum contains 2,218 windows from 385 patients, and the dynamic stratum contains 736 windows from 240 patients. Shading denotes the stable stratum. Points represent means over valid windows, and arrows indicate the preferred direction. Patients may contribute to both strata, and metric-specific valid-window counts may differ across methods.

from 9 for heart failure, and 74 from 11 for sepsis. Because these weak-label groups are small and may overlap, the analysis is exploratory and is used to characterize subgroup-specific performance patterns.

A.10.1 10-Second Forecast Horizon

Figure 8 shows that performance rankings depend on both the metric and clinical subgroup. PSSM achieves the lowest MAE and RMSE in all five groups. Our model achieves the highest correlation and R-peak F1 in four groups—stroke, ACS, heart failure, and sepsis—whereas Phase-Aligned Median Beat leads both metrics in the arrhythmia group. For example, our model obtains correlations of 0.288 and 0.239 and R-peak F1 scores of 0.523 and 0.606 in the heart-failure and sepsis groups, respectively.

Physiological summary metrics reveal additional differences. Our model has the lowest HR MAE in the stroke and heart-failure groups, while Last-10s Repeat leads in the remaining groups. Last-10s Repeat also achieves the lowest morphology distance in four groups; our model has the lowest value in ACS. These results show that temporal agreement and window-level physiological summaries produce different subgroup rankings.

A.10.2 30-Second Forecast Horizon

At 30 seconds, PSSM retains the lowest MAE and RMSE across all five groups, while periodic baselines become more competitive on temporal agreement (Figure 9). Phase-Aligned Median Beat achieves the highest R-peak F1 in every subgroup and the highest correlation in four groups; our model leads in correlation only in ACS. Our model's R-peak F1 ranges from 0.192 to 0.277, compared with 0.334 to 0.606 at 10 seconds, indicating weaker beat-timing agreement over the longer forecast horizon.

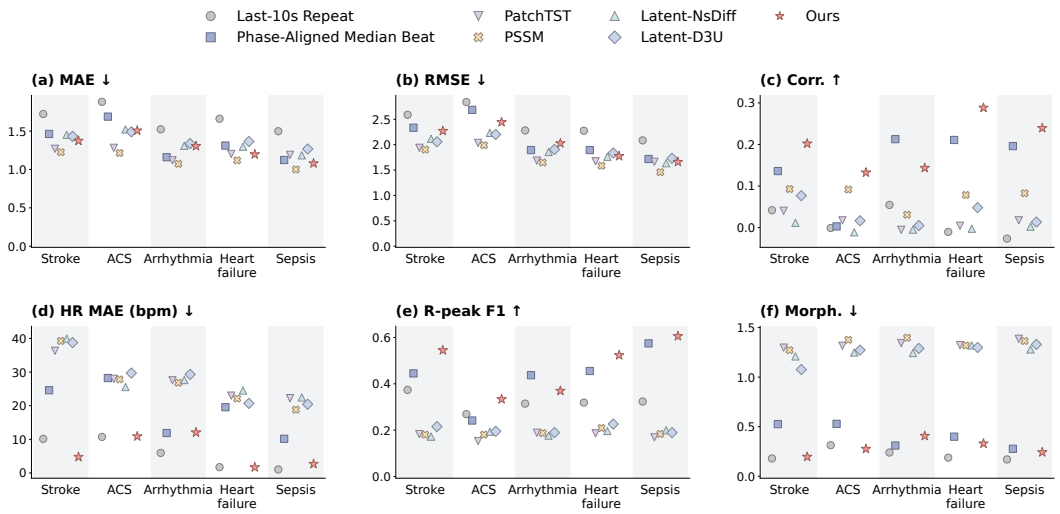


Figure 8: Exploratory six-metric evaluation within positive clinical weak-label subgroups at the 10-second forecast horizon. Stroke, ACS, arrhythmia, heart failure, and sepsis contain 7/31, 9/26, 13/63, 9/58, and 11/74 patients/windows, respectively. Points represent means over valid windows; arrows indicate the preferred direction. Subgroups overlap, and metric-specific valid window counts may differ across methods. Alternating shading separates categories and has no clinical meaning.

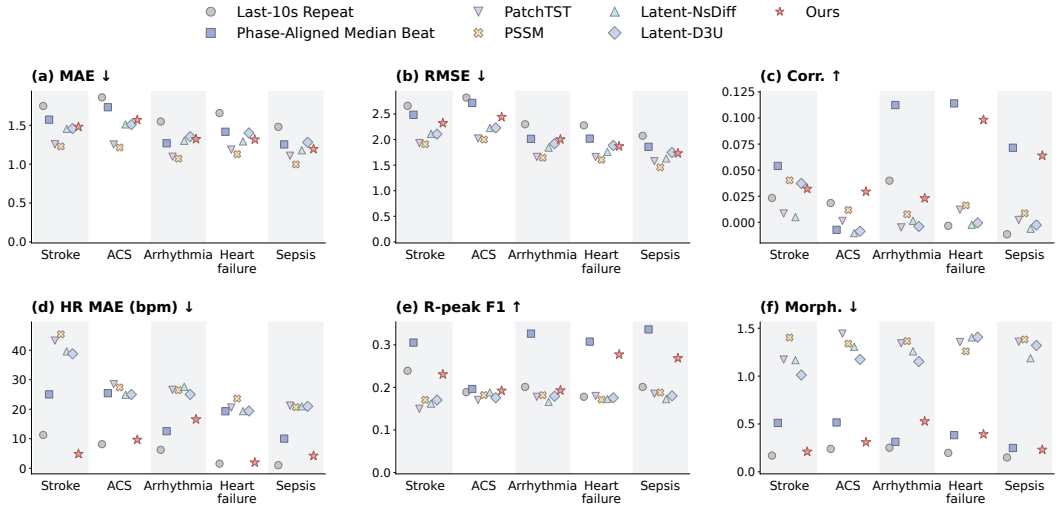


Figure 9: Exploratory six-metric evaluation within positive clinical weak-label subgroups at the 30-second forecast horizon. Subgroup patient/window counts are 7/31 for stroke, 9/26 for ACS, 13/63 for arrhythmia, 9/58 for heart failure, and 11/74 for sepsis. Points represent means over valid windows; arrows indicate the preferred direction. Subgroups overlap, and metric-specific valid window counts may differ across methods. Alternating shading is a visual aid only.

Our model nevertheless retains lower HR MAE and morphology distance than PatchTST, PSSM, Latent-NsDiff, and Latent-D3U across all five groups. It has the lowest HR MAE overall in the stroke group, whereas Last-10s Repeat leads this metric in the other groups and achieves the lowest morphology distance throughout. Within our model, the arrhythmia group has the highest HR MAE and morphology distance at both horizons, reaching 16.57 bpm and 0.528 at 30 seconds.

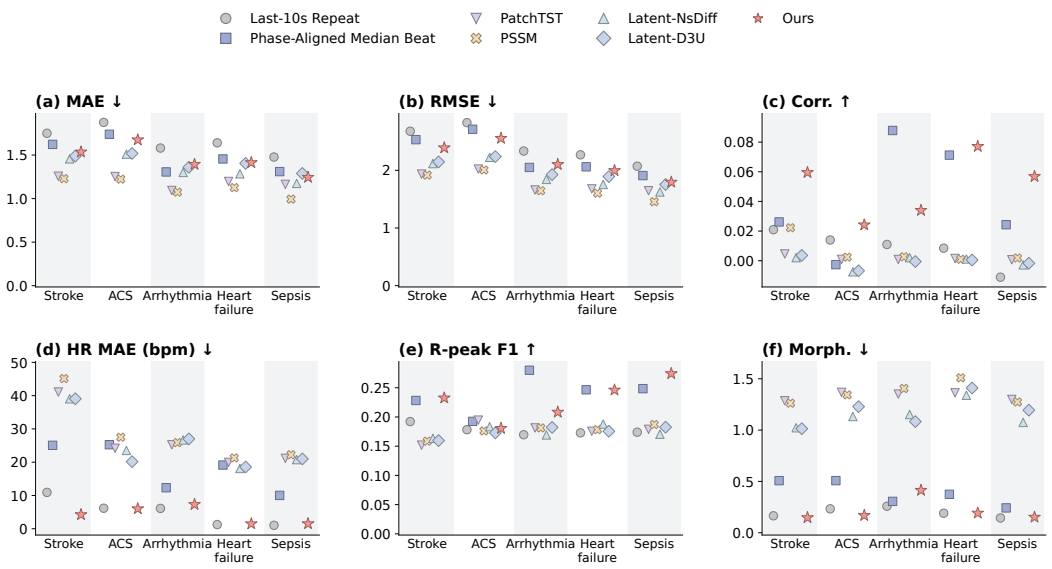


Figure 10: Exploratory six-metric comparison within positive clinical weak-label subgroups at the 60-second forecast horizon. Stroke, ACS, arrhythmia, heart failure, and sepsis contain 7/31, 9/26, 13/63, 9/58, and 11/74 patients/windows, respectively. Points represent means over valid windows; arrows indicate the preferred direction. Subgroups overlap, and metric-specific valid-window counts may differ across methods. Alternating shading separates categories and has no clinical meaning.

A.10.3 60-Second Forecast Horizon

At 60 seconds, PSSM continues to achieve the lowest MAE and RMSE across all five clinical subgroups, as shown in Figure 10. PSLD achieves the highest waveform correlation in stroke, ACS, heart failure, and sepsis, whereas Phase-Aligned Median Beat leads in the arrhythmia subgroup. R-peak F1 rankings are more heterogeneous: PSLD leads in stroke and sepsis, PatchTST in ACS, and Phase-Aligned Median Beat in arrhythmia and heart failure. In the heart-failure subgroup, PSLD and Phase-Aligned Median Beat obtain nearly identical R-peak F1 values of 0.24608 and 0.24655.

PSLD achieves lower HR MAE and morphology distance than the other learned methods in every subgroup. It also obtains the lowest values among all methods for both metrics in stroke and ACS, while Last-10s Repeat leads in arrhythmia, heart failure, and sepsis. Within PSLD, the arrhythmia subgroup has the highest HR MAE and morphology distance, at 7.30 bpm and 0.413, respectively. These results reveal subgroup-dependent differences in temporal agreement and physiological structure preservation.

A.10.4 Discussion

The clinical subgroup results reinforce the complementary roles of pointwise accuracy, temporal agreement, and physiological structure. PSSM minimizes MAE and RMSE across all five subgroups at each evaluated horizon. Our model consistently yields lower HR MAE and morphology distance than the other learned methods, while its advantages in correlation and R-peak F1 depend on the subgroup and forecast horizon. Periodic baselines remain competitive: Phase-Aligned Median Beat leads subgroup R-peak F1 throughout the 30-second evaluation and in two subgroups at 60 seconds, while Last-10s Repeat frequently achieves the lowest HR and morphology errors.

Across these weak-label subgroups, the results reveal metric-dependent differences between pointwise accuracy, temporal agreement, and physiological structure preservation.

Table 9: Waveform accuracy and physiological structure across history lengths. Bold indicates the better value within each horizon. The result has statistically significant improvements (two-sided t-test with p ¡ 0.05).

Horizon	History	MAE↓	RMSE↓	Corr.↑	HR MAE↓	R-peak F1↑	Morph.↓
10s	60s	1.442	2.222	0.172	**7.142**	**0.429**	0.272
10s	300s	**1.346**	**2.130**	**0.186**	7.198	0.422	**0.271**
30s	60s	1.437	2.244	0.043	7.426	0.223	0.310
30s	300s	**1.422**	**2.154**	**0.052**	**6.869**	**0.231**	**0.287**
60s	60s	1.480	2.227	0.021	6.681	0.188	0.269
60s	300s	**1.468**	**2.206**	**0.023**	**5.994**	**0.190**	**0.254**

Table 10: Probabilistic, distributional, and semantic metrics across history lengths. Bold indicates the better value per horizon. Coverage and width are computed from empirical 5th–95th percentile intervals using $K = 8$ trajectories, interpreted jointly, and not ranked. The result has statistically significant improvements (two-sided t-test with p ¡ 0.05).

		Probabilistic quality				ECGFounder		CLEF
Horizon	History	Energy↓	Emp. Cov. ($K = 8$)	Width	FFD↓	Cos.↑	Prob. MAE↓	Cos.↑
10s	60s	1.296	0.6405	2.706	0.203	0.860	0.0369	0.941
10s	300s	**1.287**	0.6434	2.671	**0.157**	**0.864**	**0.0351**	**0.950**
30s	60s	1.383	0.5798	2.589	1.019	0.773	0.0578	0.922
30s	300s	**1.375**	0.5966	2.626	**0.835**	**0.797**	**0.0524**	**0.933**
60s	60s	1.398	0.6016	2.781	**0.690**	0.784	0.0572	0.929
60s	300s	**1.397**	0.6022	2.735	0.695	**0.853**	**0.0411**	**0.942**

A.11 Historical Context and Conditioning Ablations

Historical context length. For each future horizon, we train separate PSLD Stage 2 models from scratch using either 60 or 300 seconds of latent ECG history. Both settings use the same 2,954 test windows from 416 patients, with eight generated waveforms per condition and 25 DDIM sampling steps. The 60-second setting retains the same waveform preprocessing protocol, including normalization based on the preceding 300-second history, while restricting the temporal latent sequence provided to Stage 2 to the final 60 seconds. Waveform, physiological, and semantic metrics evaluate the fixed first generated waveform, whereas the waveform energy score and predictive intervals use all eight trajectories. Coverage and interval width are averaged over time within each window, then over windows within each patient, and finally equally across patients.

At the 10-second horizon, the 300-second history reduces MAE from 1.442 to 1.346 and RMSE from 2.222 to 2.130, while increasing waveform correlation from 0.172 to 0.186 (Table 9). The 60-second history retains a slightly lower HR error and higher R-peak F1, whereas the two settings obtain nearly identical morphology distances. The longer history also improves the energy score, FFD, and all three semantic metrics (Table 10). Thus, longer context improves waveform and representation-level consistency at this horizon, although its effects on physiological metrics are mixed.

Longer-horizon performance. At the 30-second horizon, the 300-second history improves all waveform and physiological metrics, reducing MAE and RMSE while achieving higher correlation and R-peak F1 and lower HR and morphology errors. It also obtains a lower energy score and FFD and stronger ECGFounder and CLEF-Small consistency. These results show a broader advantage from extended context at 30 seconds.

At the 60-second horizon, the two settings remain close in pointwise error and waveform correlation, but the 300-second history performs slightly better across all waveform and physiological metrics. It also improves the ECGFounder and CLEF-Small metrics, while the 60-second history retains a slightly lower FFD (0.690 versus 0.695). The energy scores are nearly identical. The main gains from longer context at this horizon therefore appear in physiological and semantic consistency.

Predictive interval coverage. At 10 seconds, the 300-second history achieves slightly higher empirical coverage with a narrower interval. At 30 seconds, its higher coverage is accompanied by a slightly wider interval, whereas at 60 seconds the two settings obtain nearly identical coverage and the longer history produces a narrower interval. These coverage–width comparisons do not indicate a uniform effect of context length on the empirical predictive intervals.

Interpretation. The effect of historical context varies across horizons and evaluation dimensions. At 10 seconds, longer history improves pointwise accuracy, waveform correlation, and representation-level consistency, while the shorter history retains modest advantages in HR error and R-peak F1. At 30 and 60 seconds, the longer history provides more consistent improvements in waveform and physiological metrics and generally strengthens semantic consistency. Overall, extended context is most beneficial at longer horizons, although its effects on FFD and empirical prediction intervals are not uniform.

A.12 History-Relative Change Analysis

The primary benchmark metrics evaluate agreement with the observed future waveform. We additionally examine whether each method captures the direction and magnitude of physiological changes relative to the recent observed history. The analysis quantifies changes in waveform-derived physiological descriptors relative to the final 10 seconds of observed history.

Reference state and physiological descriptors. For each test window, the reference state is estimated from the final 10 seconds of observed history. We extract five descriptors from the reference segment, observed future waveform, and generated waveform: heart rate, RR interval, QRS rate, QRS amplitude, and waveform scale. For heart rate, RR interval, and QRS rate, change is defined as the difference from the reference value. For QRS amplitude and waveform scale, change is represented by the logarithm of the ratio to the reference value.

Let Δ_j and $\widehat{\Delta}_j$ denote the observed and predicted changes for descriptor j, respectively. A change is considered significant when

$$|\Delta_j| \geq \tau_j, \tag{30}$$

where the thresholds are 10 bpm for heart rate, 0.08 s for RR interval, 0.20 beats/s for QRS rate, and $\log(1.25)$ for the absolute log-ratio changes in QRS amplitude and waveform scale. The latter corresponds to a larger-to-smaller ratio of at least 1.25 in the original, non-logarithmic domain.

Change metrics. Direction accuracy measures whether the predicted and observed changes have the same sign among descriptors with a significant observed change:

$$\mathrm{DirectionAcc} = \frac{1}{|\mathcal{C}|} \sum_{j \in \mathcal{C}} \mathbb{I}\left[\mathrm{sign}(\widehat{\Delta}_j) = \mathrm{sign}(\Delta_j)\right], \tag{31}$$

where $\mathcal{C} = \{j : |\Delta_j| \geq \tau_j\}$. Magnitude error is normalized by the corresponding change threshold:

$$\mathrm{MagnitudeNMAE} = \frac{1}{|\mathcal{C}|} \sum_{j \in \mathcal{C}} \frac{|\widehat{\Delta}_j - \Delta_j|}{\tau_j}. \tag{32}$$

Change-detection recall measures the proportion of significant observed changes that are also identified in the generated waveform:

$$\mathrm{Recall} = \frac{1}{|\mathcal{C}|} \sum_{j \in \mathcal{C}} \mathbb{I}\left[|\widehat{\Delta}_j| \geq \tau_j\right]. \tag{33}$$

The false-positive rate measures how often a generated waveform crosses a change threshold when the observed future does not:

$$\mathrm{FPR} = \frac{1}{|\overline{\mathcal{C}}|} \sum_{j \in \overline{\mathcal{C}}} \mathbb{I}\left[|\widehat{\Delta}_j| \geq \tau_j\right], \tag{34}$$

where $\overline{\mathcal{C}} = \{j : |\Delta_j| < \tau_j\}$. Metrics are first averaged across eligible descriptors within each window and then summarized across test windows. For probabilistic methods, the same fixed generated waveform used by the primary waveform evaluation is used here; distributional quality is evaluated separately.

Table 11: History-relative physiological change metrics. Direction accuracy and recall are higher-is-better; magnitude NMAE and false-positive rate are lower-is-better. Bold denotes the best overall result, and underlining denotes the best learned result when a periodic baseline is best overall.

Method	Direction ↑	Magnitude ↓	Recall ↑	FPR ↓
10s horizon				
Last-10s Repeat	0.000	2.152	0.000	**0.000**
Phase-Aligned Median Beat	0.601	2.667	0.675	0.282
PatchTST	0.605	3.439	0.771	0.838
PSSM	0.597	4.261	0.756	0.807
Latent-NsDiff	0.599	2.727	**0.791**	0.846
Latent-D3U	0.589	2.664	0.746	0.817
PSLD (Ours)	**0.670**	**1.706**	0.524	0.092
30s horizon				
Last-10s Repeat	0.358	1.999	0.075	**0.003**
Phase-Aligned Median Beat	0.596	2.856	0.697	0.285
PatchTST	0.569	3.856	0.762	0.848
PSSM	0.567	4.963	**0.771**	0.858
Latent-NsDiff	0.568	2.771	0.765	0.851
Latent-D3U	0.572	2.649	0.741	0.819
PSLD (Ours)	**0.729**	**1.643**	0.576	0.415
60s horizon				
Last-10s Repeat	0.422	1.999	0.077	**0.003**
Phase-Aligned Median Beat	0.597	2.772	0.654	0.286
PatchTST	0.578	3.887	0.743	0.850
PSSM	0.571	5.628	0.742	0.864
Latent-NsDiff	0.582	2.779	**0.744**	0.849
Latent-D3U	0.585	2.630	0.728	0.821
PSLD (Ours)	**0.762**	**1.608**	0.576	0.369

Exploratory onset diagnostic. We additionally define change onset as the end of the first 5-second future window, advanced in 1-second steps, in which any descriptor crosses its threshold. A missed observed onset is assigned an error equal to the full evaluation horizon. This exploratory diagnostic measures threshold-crossing onset in waveform-derived descriptors and exhibits saturation for several learned methods.

Results. In this exploratory analysis, PSLD achieves the highest direction accuracy and the lowest normalized magnitude error at all three horizons, as shown in Table 11. Its numerical advantage in direction accuracy is largest at 30 and 60 seconds. However, its change-detection recall remains moderate, ranging from 0.524 to 0.576, and is lower than that of the other learned methods. Compared with these learned baselines, PSLD produces substantially fewer false-positive changes, suggesting a more conservative relative operating point. Its false-positive rates nevertheless increase to 0.415 and 0.369 at the 30- and 60-second horizons and exceed those of Phase-Aligned Median Beat. Last-10s Repeat produces the fewest false-positive changes because it largely preserves the recent state, but correspondingly has very low change-detection recall.

Interpretation. Overall, PSLD more accurately characterizes the direction and magnitude of ECG-derived changes when a significant change is observed, whereas its moderate recall and increased long-horizon false-positive rates indicate that identifying the occurrence of change remains more challenging. These findings indicate that PSLD captures the direction and magnitude of history-relative ECG changes more accurately, while change occurrence remains challenging, as reflected by its moderate recall and elevated long-horizon false-positive rates.

A.13 Patient-Level Paired Significance Tests

The main benchmark tables report window-level means over the test set. For statistical inference, however, we treat patients rather than individual windows as the independent units. Because all methods are evaluated on matched test windows and each patient may contribute multiple windows,

we first compute paired differences using exactly matched sample identifiers and then average these differences within each patient.

For a higher-is-better metric, the window-level difference is defined as the PSLD value minus the comparator value; for a lower-is-better metric, it is defined as the comparator value minus the PSLD value. Thus, a positive difference consistently favors PSLD. For patient i, the mean paired difference is

$$\bar{d}_i = \frac{1}{n_i} \sum_{w=1}^{n_i} d_{iw}, \tag{35}$$

where d_{iw} denotes the oriented paired difference for window w from patient i, and n_i is the number of finite paired windows contributed by that patient. The reported effect estimate is the equally weighted mean across patients:

$$\widehat{\Delta} = \frac{1}{N} \sum_{i=1}^{N} \bar{d}_i. \tag{36}$$

Because this estimand weights patients equally, $\widehat{\Delta}$ need not equal the direct difference between the window-level means reported in the main benchmark tables.

We obtain marginal 95% confidence intervals using 10,000 patient-level bootstrap resamples and compute two-sided patient-level sign-flip randomization tests using 100,000 permutations. Holm correction is applied once across a prespecified family of 108 comparisons, comprising PSLD against six baselines, six primary waveform and physiological metrics, and three future horizons. The confidence intervals are unadjusted percentile bootstrap intervals, whereas statistical significance is determined using the Holm-adjusted p values. A confidence interval may therefore exclude zero even when the corresponding comparison is not significant after multiplicity correction.

History-relative change metrics and clinical weak-label subgroup analyses are treated as exploratory and are excluded from this confirmatory family. The analysis requires identical sample sets for each method pair and does not use intersection-based filtering. Up to 2,954 matched windows from 416 patients are available for each comparison; metric-specific effective counts are reported with the table.

For readability, Table 12 presents selected comparisons with the two periodic references and PSSM, the strongest deterministic baseline in pointwise accuracy. An asterisk indicates a Holm-adjusted $p < 0.05$. Complete effect estimates, confidence intervals, adjusted p values, and sample counts for all 108 comparisons are provided in the supplementary statistical report.

Comparison with PSSM. PSSM achieves significantly lower MAE than PSLD at all three horizons, confirming its advantage in pointwise waveform accuracy. PSLD achieves significantly higher waveform correlation and R-peak F1 and significantly lower HR MAE and morphology distance throughout. The paired results reveal complementary strengths in pointwise accuracy, temporal agreement, and physiological structure preservation.

Comparison with periodic baselines. PSLD achieves significantly lower MAE than both periodic baselines at every horizon. Relative to Last-10s Repeat, PSLD also achieves significantly higher waveform correlation and R-peak F1 throughout. Last-10s Repeat nevertheless retains significantly lower morphology distance at all three horizons and lower HR MAE at 10 and 30 seconds. Their HR MAE difference at 60 seconds is not significant after Holm correction.

Relative to Phase-Aligned Median Beat, PSLD achieves significantly lower HR MAE and morphology distance at every horizon. PSLD has significantly higher waveform correlation and R-peak F1 at 10 seconds. At 30 and 60 seconds, Phase-Aligned Median Beat achieves significantly higher R-peak F1, whereas the correlation differences are not statistically significant.

Summary. Three comparisons displayed in Table 12 do not reach the Holm-corrected significance threshold: correlation between PSLD and Phase-Aligned Median Beat at 30 and 60 seconds, and HR MAE between PSLD and Last-10s Repeat at 60 seconds. In the complete 108-comparison family, the 30-second MAE comparison between PSLD and Latent-D3U is also not significant, yielding four nonsignificant comparisons in total.

Table 12: Selected patient-level paired comparisons across future horizons. Positive $\widehat{\Delta}$ favors PSLD. * indicates Holm-adjusted $p < 0.05$ across the complete family of 108 comparisons.

Metric	10s	30s	60s
PSLD vs. Last-10s Repeat			
MAE	$0.363\,[0.338, 0.389]^*$	$0.297\,[0.284, 0.310]^*$	$0.253\,[0.244, 0.263]^*$
Correlation	$0.171\,[0.153, 0.189]^*$	$0.046\,[0.038, 0.054]^*$	$0.017\,[0.012, 0.022]^*$
HR MAE	$-1.229\,[-1.837, -0.635]^*$	$-1.604\,[-2.314, -0.909]^*$	$-0.687\,[-1.375, -0.036]$
R-peak F1	$0.237\,[0.216, 0.258]^*$	$0.062\,[0.052, 0.073]^*$	$0.024\,[0.018, 0.030]^*$
Morphology	$-0.087\,[-0.100, -0.075]^*$	$-0.122\,[-0.137, -0.106]^*$	$-0.096\,[-0.112, -0.079]^*$
PSLD vs. Phase-Aligned Median Beat			
MAE	$0.064\,[0.044, 0.085]^*$	$0.092\,[0.078, 0.107]^*$	$0.087\,[0.075, 0.100]^*$
Correlation	$0.059\,[0.046, 0.071]^*$	$-0.002\,[-0.010, 0.005]$	$-0.005\,[-0.011, -0.00005]$
HR MAE	$15.366\,[13.130, 17.718]^*$	$15.610\,[13.259, 18.044]^*$	$16.678\,[14.293, 19.160]^*$
R-peak F1	$0.073\,[0.058, 0.088]^*$	$-0.016\,[-0.026, -0.007]^*$	$-0.018\,[-0.025, -0.012]^*$
Morphology	$0.162\,[0.131, 0.194]^*$	$0.130\,[0.097, 0.164]^*$	$0.157\,[0.124, 0.191]^*$
PSLD vs. PSSM			
MAE	$-0.205\,[-0.226, -0.183]^*$	$-0.257\,[-0.267, -0.248]^*$	$-0.303\,[-0.312, -0.294]^*$
Correlation	$0.072\,[0.058, 0.086]^*$	$0.032\,[0.026, 0.038]^*$	$0.018\,[0.014, 0.022]^*$
HR MAE	$23.225\,[21.670, 24.755]^*$	$24.300\,[22.727, 25.841]^*$	$24.996\,[23.420, 26.568]^*$
R-peak F1	$0.243\,[0.224, 0.261]^*$	$0.051\,[0.043, 0.059]^*$	$0.016\,[0.011, 0.020]^*$
Morphology	$1.078\,[1.053, 1.104]^*$	$1.057\,[1.029, 1.085]^*$	$1.071\,[1.042, 1.099]^*$

Effects are equal-patient mean paired differences with marginal patient-bootstrap 95% confidence intervals. All displayed comparisons use 416 patients and 2,954 paired windows, except HR MAE and morphology against Phase-Aligned Median Beat, which use 416 patients and 2,950 windows. Exact Holm-adjusted *p* values for the complete 108-comparison family are provided in the supplementary statistical report.

Table 13: Stage-1 physiological supervision ablation. Bold indicates the better value at the reported precision. The result has statistically significant improvements (two-sided t-test with p ¡ 0.05).

Metric	Without supervision	With supervision
MAE ($\times 10^{-3}$)↓	8.627	**8.598**
RMSE ($\times 10^{-3}$)↓	**16.077**	16.663
Correlation↑	1.000	1.000
SNR (dB)↑	**42.800**	42.480
R-peak F1↑	0.996	0.996
R-peak timing MAE (ms)↓	0.650	**0.610**
Median RR error (ms)↓	5.257	**4.638**
Physiological state MAE↓	0.368	**0.091**

Overall, the patient-level paired analysis identifies complementary strengths across the evaluated paradigms. PSLD leads PSSM in temporal agreement and physiological fidelity, deterministic prediction retains an advantage in pointwise accuracy, and periodic continuation remains competitive in recent-beat structure preservation.

A.14 Physiological Supervision Ablation

A.14.1 Stage 1: Reconstruction and Physiological State Prediction

We compare Stage-1 VAEs trained with and without physiological supervision on the same test set of 2,929 ECG windows. Both models use 48 latent dimensions and a latent rate of 10 tokens per second. We evaluate waveform reconstruction, preservation of beat timing, and normalized errors of the physiological state outputs.

Reconstruction fidelity. Both models achieve high reconstruction fidelity, with waveform correlations above 0.99995 and R-peak F1 scores above 0.9958 (Table 13). Physiological supervision

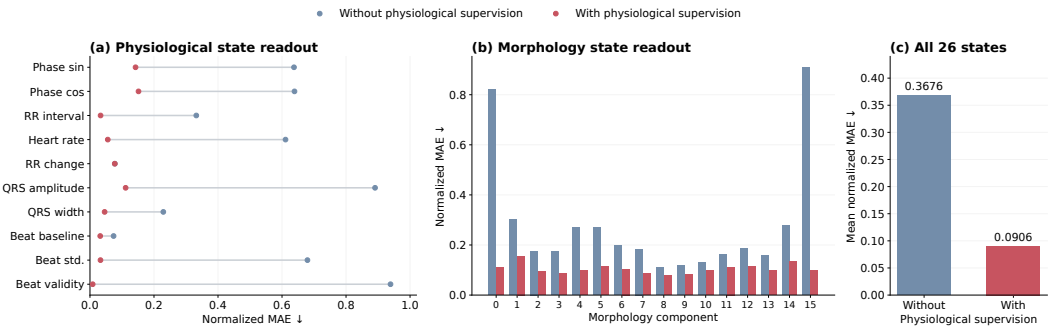


Figure 11: Physiological state prediction errors with and without Stage-1 physiological supervision. Lower values are preferred.

produces slightly lower waveform MAE and slightly higher R-peak recall and F1, while the model without supervision achieves lower RMSE and higher reconstruction SNR. R-peak timing MAE decreases from 0.650 to 0.610 ms, and median RR error decreases from 5.257 to 4.638 ms. These mixed results indicate that physiological supervision maintains high reconstruction quality without uniformly improving every reconstruction metric.

Physiological state prediction. The clearest difference occurs in physiological state prediction (Figure 11). Mean normalized MAE across the 26 state components decreases from 0.3676 to 0.0906, a relative reduction of 75.4%. RR-interval error decreases from 0.3322 to 0.0331, HR error from 0.6109 to 0.0556, and QRS-amplitude error from 0.8906 to 0.1114. Phase, QRS width, beat baseline, and beat variability also show lower errors, as do all 16 morphology components. In contrast, RR-change error remains nearly unchanged, at 0.07760 and 0.07744. Thus, the improvement is broad but varies substantially across physiological quantities.

Qualitative reconstruction comparison. We additionally compare matched reconstructions of the same test windows using full-window overlays, a shared local zoom, and absolute reconstruction-error curves, as Figure 12. All methods use the same target waveform and display interval, without post-hoc temporal alignment or amplitude rescaling.

Interpretation. The Stage-1 results characterize waveform reconstruction and physiological state prediction as complementary properties of the latent representation. Physiological supervision substantially improves the designated state readout while preserving near-perfect waveform reconstruction. The subsequent Stage-2 comparison directly evaluates how this supervision affects future waveform generation.

A.14.2 Stage 2: Downstream ECG Forecasting

We compare downstream forecasting models trained using Stage-1 representations learned with and without physiological supervision. Both settings use 300-second ECG histories and are evaluated on the same 2,954 test windows from 416 patients at forecast horizons of 10, 30, and 60 seconds. Waveform, physiological, and semantic metrics evaluate the fixed first generated waveform, while the waveform energy score and predictive intervals use all eight generated waveforms per condition. Coverage and interval width are averaged over time within each window, then over windows within each patient, and finally equally across patients.

10s horizon. At the 10-second horizon, physiological supervision improves all reported waveform and physiological metrics (Table 14). It reduces MAE from 1.359 to 1.346 and RMSE from 2.149 to 2.130, while increasing correlation from 0.182 to 0.186 and R-peak F1 from 0.411 to 0.422. HR MAE and morphology distance also decrease. The supervised representation further improves the energy score, FFD, and all three semantic metrics (Table 15).

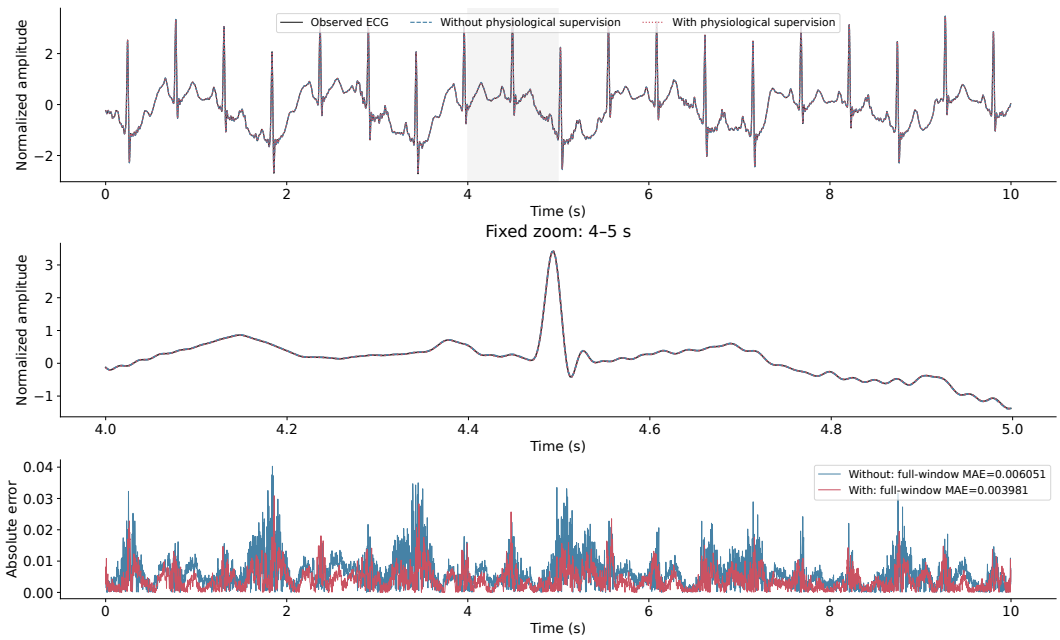


Figure 12: Stage 1 VAE reconstruction visualization.

Table 14: Downstream waveform accuracy and physiological structure. Bold indicates the better value within each horizon at the reported precision. The result has statistically significant improvements (two-sided t-test with p ¡ 0.05).

Horizon	Supervision	MAE↓	RMSE↓	Corr.↑	HR MAE↓	R-peak F1↑	Morph.↓
10s	Without	1.359	2.149	0.182	7.847	0.411	0.275
10s	With	**1.346**	**2.130**	**0.186**	**7.198**	**0.422**	**0.271**
30s	Without	1.450	2.189	0.049	6.980	0.227	0.292
30s	With	**1.422**	**2.154**	**0.052**	**6.869**	**0.231**	**0.287**
60s	Without	1.472	2.212	0.021	6.198	0.189	0.259
60s	With	**1.468**	**2.206**	**0.023**	**5.994**	**0.190**	**0.254**

30s horizon. At the 30-second horizon, physiological supervision again improves all waveform and physiological metrics. MAE and RMSE decrease from 1.450 and 2.189 to 1.422 and 2.154, respectively, while correlation and R-peak F1 increase. HR MAE decreases from 6.980 to 6.869 bpm, and morphology distance decreases from 0.292 to 0.287. Supervision also yields a lower energy score and FFD and stronger consistency across ECGFounder and CLEF-Small.

60s horizon. At the 60-second horizon, physiological supervision maintains consistent improvements across all waveform and physiological metrics. It slightly reduces MAE from 1.472 to 1.468 and RMSE from 2.212 to 2.206, while increasing correlation from 0.021 to 0.023. HR MAE and morphology distance decrease from 6.198 to 5.994 bpm and from 0.259 to 0.254, respectively, while R-peak F1 increases slightly. Supervision also improves the energy score, FFD, and all three semantic metrics.

Predictive interval coverage. Physiological supervision produces slightly higher empirical coverage with narrower intervals at all three horizons. Coverage changes from 0.6403 to 0.6434 at 10 seconds, from 0.5859 to 0.5966 at 30 seconds, and from 0.6002 to 0.6022 at 60 seconds. The corresponding reduction in interval width is largest at 30 seconds.

Table 15: Downstream probabilistic, distributional, and semantic metrics. Bold indicates the better value per horizon. Coverage and width are computed from empirical 5th-95th percentile intervals using $K = 8$ trajectories, interpreted jointly, and not ranked. The result has statistically significant improvements (two-sided t-test with p ¡ 0.05).

		Probabilistic quality				ECGFounder		CLEF
Horizon	Supervision	Energy↓	Emp. Cov. ($K = 8$)	Width	FFD↓	Cos.↑	Prob. MAE↓	Cos.↑
10s	Without	1.289	0.6403	2.695	0.186	0.860	0.0368	0.949
10s	With	**1.287**	0.6434	2.671	**0.157**	**0.864**	**0.0351**	**0.950**
30s	Without	1.382	0.5859	2.802	0.920	0.793	0.0563	0.929
30s	With	**1.375**	0.5966	2.626	**0.835**	**0.797**	**0.0524**	**0.933**
60s	Without	1.405	0.6002	2.757	0.736	0.801	0.0524	0.933
60s	With	**1.397**	0.6022	2.735	**0.695**	**0.853**	**0.0411**	**0.942**

Interpretation. Stage-1 physiological supervision yields consistent, although generally modest, improvements in waveform accuracy and physiological structure across all three horizons. It also improves the energy score, FFD, and semantic consistency throughout. The gains in pointwise accuracy are small, particularly at 60 seconds, whereas the improvements in physiological and representation-level metrics are more consistent. These results support the value of physiological supervision across complementary evaluation dimensions.

A.15 Ablation Study on Multiscale Historical Conditioning

We examine the contributions of multiscale history C_{multi} and clean recent-prefix tokens P_{recent} to Stage 2 generation. We construct three ablated variants by disabling either component or both, and compare them with full PSLD in Tables 16 and 17. Disabling C_{multi} replaces multiscale aggregation with single-scale conditioning over the same 300-second history, rather than removing historical information. Disabling P_{recent} removes the direct clean-prefix token pathway while retaining historical and rhythm/phase conditioning. All configurations reuse the frozen Stage 1 VAE and are trained separately in Stage 2 with the same DiT depth and width.

All configurations are evaluated on the same 2,954 test windows at each horizon using EMA checkpoints, 25 DDIM sampling steps, a guidance scale of 1.0, and eight generated trajectories per condition. Primary waveform and physiological measurements use the fixed first generated sample, and physiological evaluation follows the original benchmark detector. The ablations hold the latent representation and DiT depth and width fixed while varying the conditioning configuration.

Effect of clean-prefix conditioning. The clean prefix consistently improves waveform accuracy and physiological structure under both historical representations. With multiscale history retained, adding the prefix increases 10-second correlation from 0.105 to 0.186 and R-peak F1 from 0.321 to 0.422, while reducing heart-rate MAE from 10.780 to 7.198 and morphology distance from 0.413 to 0.271. Morphology also improves at 30 and 60 s, from 0.351 to 0.287 and from 0.337 to 0.254, respectively. Under single-scale history, adding the prefix likewise improves all six waveform and physiological measurements at every horizon. These observations are consistent with recent-prefix information supporting short-horizon temporal alignment and beat-structure preservation.

Effect of multiscale historical conditioning. With the clean prefix retained, multiscale conditioning further improves all six waveform and physiological measurements across the three horizons. For example, morphology distance decreases from 0.295 to 0.271 at 10 s, from 0.309 to 0.287 at 30 s, and from 0.287 to 0.254 at 60 s. Without the prefix, multiscale history also improves most measurements, although 30-second correlation decreases from 0.042 to 0.028. Thus, multiscale aggregation provides the most consistent improvements when combined with clean-prefix conditioning, rather than improving every measurement in isolation.

Probabilistic, distributional, and semantic consistency. Full PSLD achieves the lowest energy score and FFD, the highest ECGFounder and CLEF cosine similarities, and the lowest ECGFounder probability MAE at every horizon. With the prefix retained, adding multiscale history reduces FFD

Table 16: Conditioning ablations for waveform accuracy and physiological structure. Both components enabled correspond to full PSLD. Bold denotes the best result per horizon.

Components		Waveform accuracy			Physiological structure		
C_{multi}	P_{recent}	MAE ↓	RMSE ↓	Corr. ↑	HR MAE ↓	R-peak F1 ↑	Morph. ↓
10-second future horizon							
–	–	1.406	2.162	0.084	11.160	0.299	0.451
✓	–	1.399	2.148	0.105	10.780	0.321	0.413
–	✓	1.357	2.137	0.181	7.675	0.401	0.295
✓	✓	**1.346**	**2.130**	**0.186**	**7.198**	**0.422**	**0.271**
30-second future horizon							
–	–	1.449	2.170	0.042	8.338	0.208	0.373
✓	–	1.430	2.162	0.028	7.345	0.217	0.351
–	✓	1.428	2.158	0.049	6.953	0.224	0.309
✓	✓	**1.422**	**2.154**	**0.052**	**6.869**	**0.231**	**0.287**
60-second future horizon							
–	–	1.483	2.254	0.017	6.777	0.180	0.370
✓	–	1.476	2.233	0.019	6.313	0.184	0.337
–	✓	1.473	2.218	0.022	6.016	0.188	0.287
✓	✓	**1.468**	**2.206**	**0.023**	**5.994**	**0.190**	**0.254**

Table 17: Probabilistic, distributional, and semantic evaluation of conditioning ablations. Bold denotes the best result per horizon. Coverage and width are interpreted jointly and are not ranked.

Components		Probabilistic quality			Distribution	ECGFounder		CLEF
C_{multi}	P_{recent}	Energy ↓	Cov. (5–95%)	Width	FFD ↓	Cos. ↑	Prob. MAE ↓	Cos. ↑
10-second future horizon								
–	–	1.355	0.5968	2.536	1.017	0.737	0.0597	0.907
✓	–	1.349	0.5936	2.566	1.005	0.752	0.0556	0.918
–	✓	1.288	0.6379	2.629	0.205	0.850	0.0357	0.940
✓	✓	**1.287**	0.6434	2.671	**0.157**	**0.864**	**0.0351**	**0.950**
30-second future horizon								
–	–	1.392	0.5915	2.645	1.115	0.748	0.0676	0.909
✓	–	1.385	0.5795	2.541	1.012	0.761	0.0605	0.914
–	✓	1.379	0.5974	2.621	0.854	0.785	0.0577	0.925
✓	✓	**1.375**	0.5966	2.626	**0.835**	**0.797**	**0.0524**	**0.933**
60-second future horizon								
–	–	1.407	0.5934	2.695	1.113	0.761	0.0568	0.914
✓	–	1.406	0.5881	2.666	1.037	0.739	0.0535	0.923
–	✓	1.401	0.5987	2.714	0.785	0.798	0.0531	0.935
✓	✓	**1.397**	0.6022	2.735	**0.695**	**0.853**	**0.0411**	**0.942**

from 0.205 to 0.157 at 10 s, from 0.854 to 0.835 at 30 s, and from 0.785 to 0.695 at 60 s. At 60 s, it also increases ECGFounder cosine similarity from 0.798 to 0.853 and reduces probability MAE from 0.0531 to 0.0411. However, multiscale conditioning without a prefix does not uniformly improve semantic measurements: its 60-second ECGFounder cosine similarity is lower than that of the single-scale configuration. Coverage and interval width also vary non-monotonically and are interpreted jointly rather than ranked independently.

Discussion. Full PSLD achieves the best point estimates on all ranked metrics across the three horizons. Clean-prefix conditioning contributes the larger improvements in short-horizon temporal

alignment, while multiscale history provides additional gains when the prefix is retained. For example, adding multiscale history to the prefix-equipped model reduces morphology distance from 0.295 to 0.271 at 10 seconds and from 0.287 to 0.254 at 60 seconds. Together, these results support combining recent high-resolution context with multiscale historical information.

A.16 Evaluation Validity and Statistical Reliability

A.16.1 Evaluation Validity.

We use patient-disjoint data splits to ensure that windows from the same patient do not appear across the training, validation, and test sets. All methods are evaluated on matched test windows against the preprocessed observed ECG signals. Waveform accuracy, physiological structure, and semantic consistency metrics are computed using a fixed generated waveform (draw index 0), whereas energy scores and prediction intervals use all eight trajectories generated for each historical condition. This protocol separately evaluates individual-trajectory quality and probabilistic modeling performance.

Stable and dynamic windows are defined using uniform, prespecified thresholds for physiological changes between the observed history and future. Clinical weak labels are used only for evaluation stratification and are not provided to the model as future information. These settings enable comparisons under consistent evaluation conditions while distinguishing the complementary forecasting capabilities reflected by pointwise error, temporal alignment, physiological structure, and semantic consistency.

A.16.2 Finite-Sample Prediction-Interval Coverage

For each historical condition, we estimate pointwise prediction intervals using the empirical 5th and 95th percentiles of $K = 8$ generated waveforms, with linear interpolation between adjacent order statistics. Coverage and interval width are averaged over future time points within each window, then across windows within each patient, and finally equally across patients. These intervals assess pointwise rather than simultaneous trajectory coverage.

With only eight generated waveforms, empirical tail-quantile intervals exhibit substantial finite-sample effects. Even when the observation and generated waveforms are conditionally i.i.d. draws from the same continuous distribution, the expected coverage of linearly interpolated sample quantiles depends on both the predictive distribution and the quantile estimator. Linear interpolation in the waveform-value domain cannot generally be interpreted as interpolation between cumulative probabilities. Therefore, the empirical 5th–95th percentile interval does not have a distribution-independent expected coverage of 0.70.

A distribution-free result is available for the interval spanning the sample minimum and maximum. For K i.i.d. samples from a continuous distribution, the probability that an independent observation falls between the two extrema is

$$\mathbb{P}\big(X_{(1)} \le X_{\text{new}} \le X_{(K)}\big) = \frac{K-1}{K+1}. \tag{37}$$

For $K = 8$, this probability is $7/9 \approx 0.778$. Because the linearly interpolated 5th–95th percentile interval lies within the sample range, its expected coverage is no greater than this value, while its exact expectation remains distribution dependent.

PSLD obtains empirical coverage of 0.6434, 0.5966, and 0.6022 at the 10-, 30-, and 60-second horizons, respectively. At 30 and 60 seconds, Latent-D3U achieves higher empirical coverage with narrower intervals. Coverage and interval width characterize the empirical intervals produced under the available sampling budget, while the energy score evaluates the sampled predictive distribution without requiring tail-quantile estimation.

A.17 Robustness to Multi-Sample Evaluation

Evaluation protocol. We assess the sensitivity of trajectory-level evaluation to the number of generated samples using PSLD, Latent-NsDiff, and Latent-D3U at forecast horizons of 10, 30, and 60 seconds. All settings use the same 2,954 test windows. For each window, we compute waveform MAE, RMSE, correlation, heart-rate MAE, R-peak F1, and morphology distance separately for each

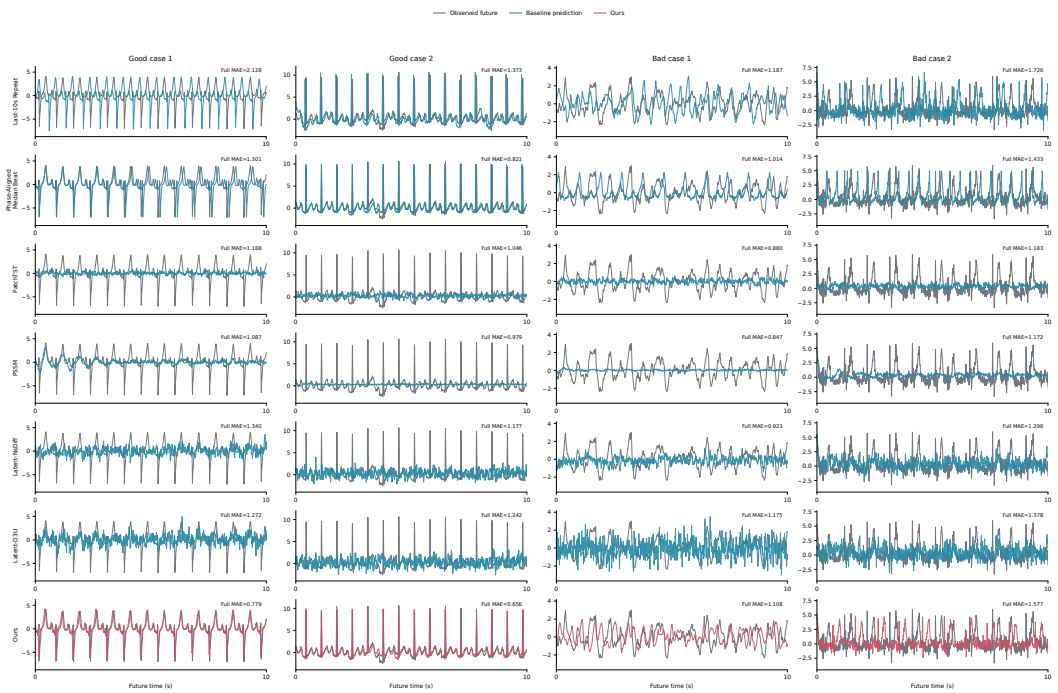


Figure 13: Qualitative examples and failure cases at the 10-second forecast horizon. Rows correspond to seven methods, and columns contain two favorable examples and two failure cases selected according to our model's composite performance. Gray denotes the observed future, blue baseline predictions, and red our prediction. Complete 10-second forecasts are displayed. Predictions use fixed draw index 0, amplitude limits are shared within each column, and labels report full-horizon MAE.

generated trajectory. We then average each metric across trajectories within the window and equally across test windows:

$$\widehat{M}_K = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{K}\sum_{k=1}^{K} m\left(\hat{x}_i^{(k)}, x_i\right)\right], \tag{38}$$

where x_i denotes the observed future and $\hat{x}_i^{(k)}$ denotes the kth generated trajectory. We evaluate $K \in \{1, 8, 16, 32, 64\}$, with $K = 1$ corresponding to fixed draw 0. This procedure averages per-trajectory metrics rather than averaging waveforms before evaluation or selecting the best trajectory. All reported metrics have valid values for every evaluated window and draw.

The tables report values recomputed within this experiment, with a fixed checkpoint for each model and horizon across sample counts. They are distinguished from the main tables, whose trajectory-level metrics evaluate fixed draw 0 of an eight-sample generation run. In particular, the PSLD 60-second experiment here uses the best checkpoint, whereas the main 60-second results use the final checkpoint. Comparisons of sample counts are therefore made within this experiment.

Single-trajectory versus multi-sample evaluation. Moving from fixed draw 0 to averaging over multiple trajectories produces small changes in the reported metrics without reversing the principal physiological comparisons. At the 10-second horizon, PSLD's MAE changes from 1.346 at $K = 1$ to 1.349 at $K = 8$, while correlation changes from 0.186 to 0.183 and R-peak F1 from 0.422 to 0.419 (Table 18). Heart-rate MAE decreases from 7.205 to 7.129 bpm. At 30 and 60 seconds, the corresponding changes are also small (Tables 19 and 20). The results therefore suggest that the fixed trajectory captures the broad performance pattern, although its metric values need not equal the expected performance across generated trajectories.

Table 18: Single-trajectory and multi-sample mean metrics at the 10-second horizon.

Method	K	MAE↓	RMSE↓	Corr.↑	HR MAE↓	R-peak F1↑	Morph.↓
Latent-NsDiff	1	1.397	1.983	0.004	30.255	0.174	1.307
	8	1.396	1.982	0.005	30.553	0.175	1.311
	16	1.396	1.982	0.005	30.509	0.175	1.313
	32	1.396	1.982	0.005	30.477	0.175	1.312
	64	1.396	1.982	0.005	30.468	0.175	1.313
Latent-D3U	1	1.408	2.001	0.031	30.176	0.178	1.312
	8	1.407	1.999	0.032	30.424	0.178	1.304
	16	1.408	1.999	0.032	30.374	0.178	1.307
	32	1.408	1.999	0.032	30.300	0.178	1.306
	64	1.408	1.999	0.032	30.258	0.178	1.305
PSLD (Ours)	1	1.346	2.130	0.186	7.205	0.422	0.271
	8	1.349	2.135	0.183	7.129	0.419	0.271
	16	1.350	2.136	0.182	7.149	0.418	0.270
	32	1.351	2.137	0.182	7.118	0.418	0.269
	64	1.350	2.136	0.182	7.101	0.418	0.270

Table 19: Single-trajectory and multi-sample mean metrics at the 30-second horizon.

Method	K	MAE↓	RMSE↓	Corr.↑	HR MAE↓	R-peak F1↑	Morph.↓
Latent-NsDiff	1	1.401	1.989	0.001	29.326	0.169	1.260
	8	1.401	1.989	0.001	29.111	0.170	1.258
	16	1.401	1.989	0.001	29.048	0.169	1.254
	32	1.401	1.989	0.001	29.037	0.170	1.253
	64	1.401	1.989	0.001	28.995	0.170	1.252
Latent-D3U	1	1.436	2.033	0.004	29.043	0.169	1.265
	8	1.436	2.032	0.004	29.130	0.171	1.259
	16	1.436	2.033	0.004	29.151	0.171	1.260
	32	1.436	2.033	0.004	29.160	0.171	1.258
	64	1.436	2.033	0.004	29.124	0.171	1.258
PSLD (Ours)	1	1.422	2.154	0.052	6.869	0.231	0.287
	8	1.425	2.157	0.051	6.809	0.230	0.287
	16	1.425	2.157	0.051	6.821	0.230	0.287
	32	1.425	2.157	0.051	6.814	0.230	0.288
	64	1.426	2.157	0.051	6.799	0.230	0.287

Sensitivity to sample count. Increasing K from 8 to 64 has a limited effect on the multi-sample mean metrics. For PSLD at 10 seconds, MAE changes from 1.3493 to 1.3505, correlation from 0.1829 to 0.1823, and R-peak F1 from 0.4190 to 0.4184. At 30 seconds, MAE changes from 1.4252 to 1.4255 and R-peak F1 from 0.2301 to 0.2302. At 60 seconds, MAE changes from 1.4673 to 1.4672 and R-peak F1 from 0.1911 to 0.1915. Latent-NsDiff and Latent-D3U likewise show small changes across sample counts. These changes are not consistently monotonic: increasing the number of evaluated trajectories refines the estimate of average trajectory quality rather than directly improving individual predictions.

Consistency of method comparisons. Across all three forecast horizons and all five sample counts, PSLD achieves higher waveform correlation and R-peak F1, together with lower heart-rate MAE and morphology distance, than Latent-NsDiff and Latent-D3U. Pointwise accuracy remains metric- and horizon-dependent. PSLD achieves lower MAE than both baselines at 10 seconds, falls between the two baselines at 30 seconds, and has higher MAE than both at 60 seconds. Its RMSE is higher than both baselines at every horizon. Thus, the distinction between pointwise accuracy and physiological structure persists when evaluation is averaged over multiple generated trajectories.

Interpretation. The main physiological comparison is robust to evaluating multiple trajectories: PSLD's advantages over the two latent diffusion baselines in temporal agreement and physiological

Table 20: Single-trajectory and multi-sample mean metrics at the 60-second horizon.

Method	K	MAE↓	RMSE↓	Corr.↑	HR MAE↓	R-peak F1↑	Morph.↓
Latent-NsDiff	1	1.401	1.992	0.000	28.615	0.170	1.194
	8	1.401	1.992	0.000	28.546	0.170	1.205
	16	1.401	1.992	0.000	28.555	0.170	1.203
	32	1.401	1.992	0.000	28.536	0.170	1.201
	64	1.401	1.992	0.000	28.530	0.170	1.201
Latent-D3U	1	1.447	2.046	0.000	28.799	0.169	1.219
	8	1.447	2.046	0.000	28.861	0.170	1.218
	16	1.447	2.046	0.000	28.854	0.170	1.220
	32	1.447	2.046	0.000	28.846	0.170	1.217
	64	1.447	2.046	0.000	28.843	0.170	1.217
PSLD (Ours)	1	1.468	2.206	0.023	5.996	0.190	0.254
	8	1.467	2.206	0.024	6.061	0.191	0.254
	16	1.467	2.206	0.024	6.021	0.192	0.255
	32	1.467	2.206	0.024	6.010	0.192	0.255
	64	1.467	2.206	0.024	6.001	0.191	0.255

structure are not confined to a single fixed draw. The small changes from $K = 8$ to $K = 64$ indicate limited empirical sensitivity of the reported mean metrics to sample count under the evaluated settings, rather than establishing a universal convergence threshold.

A.18 Qualitative Examples and Failure Cases

Case selection and visualization protocol. We compare seven methods on four matched examples at the 10-, 30-, and 60-second forecast horizons. Rows correspond to methods, and columns correspond to cases, including two favorable examples and two failure cases. Cases are selected independently at each horizon, and the four cases within each figure come from different patients.

Within each column, all methods use the same sample, and probabilistic models display fixed draw index 0. Gray denotes the observed future, blue denotes baseline predictions, and red denotes our prediction. Amplitude limits are shared across methods within each case, without post-hoc temporal alignment or amplitude rescaling. The 10-second figure displays complete forecasts; the 30-second figure shows 0–5, 12.5–17.5, and 25–30 seconds; and the 60-second figure shows 0–5, 27.5–32.5, and 55–60 seconds. Light blue, light orange, and light green backgrounds distinguish the beginning, middle, and end segments, respectively, while double slashes indicate omitted intervals. Case selection and annotated MAEs use the complete forecast horizon, including intervals omitted from display.

10s horizon. As shown in Figure 13, in the two favorable examples, our model reproduces prominent cardiac complexes and exhibits closer temporal agreement than most baselines. Its MAEs are 0.779 and 0.656, compared with 1.087 and 0.979 for PSSM. PatchTST and PSSM exhibit pronounced amplitude attenuation, particularly in the second example, despite relatively low pointwise errors. Periodic continuation preserves sharp cardiac complexes but places some peaks at incorrect times, while the two latent diffusion baselines reproduce beat structure less faithfully in these examples.

The failure cases reveal limitations of our model. In the first, the generated oscillations do not closely follow the observed waveform; in the second, prominent generated peaks are temporally misaligned, and the inter-beat signal also differs from the observed future. Although PSSM produces strongly attenuated predictions, it achieves lower MAE in both cases. These examples show that retaining visible cardiac complexes is insufficient unless their timing and morphology also match the observed future.

30s horizon. As shown in Figure 14, in the favorable examples, our model retains prominent cardiac complexes in the beginning, middle, and end segments, although local timing and amplitude discrepancies remain. The second favorable example also shows that a favorable composite ranking

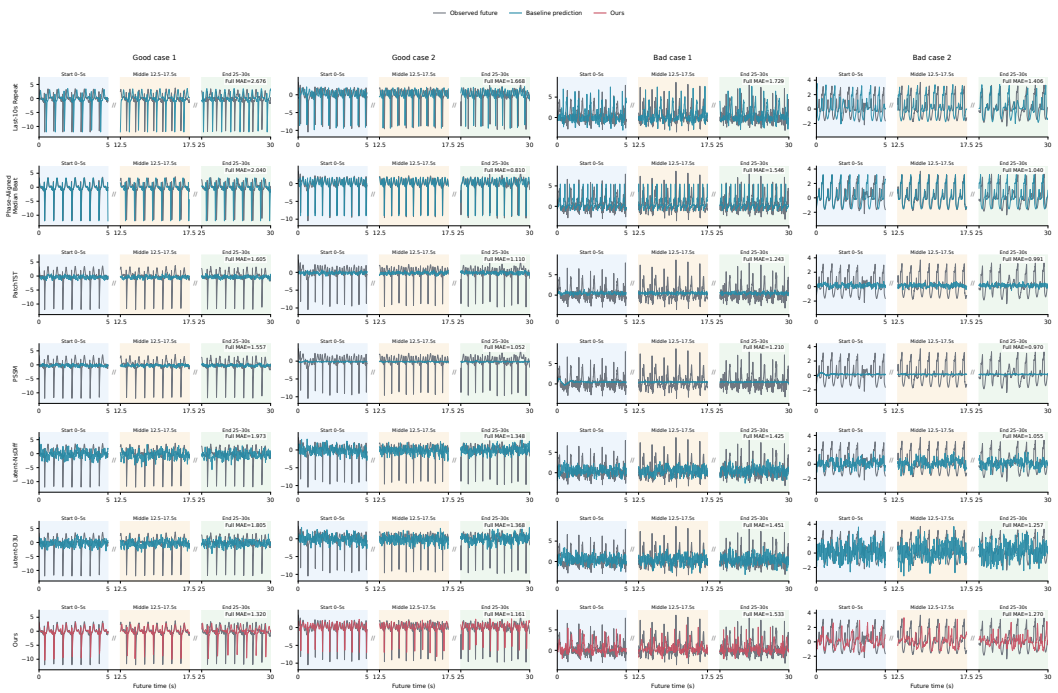


Figure 14: Qualitative examples and failure cases at the 30-second forecast horizon. Rows denote methods and columns denote matched cases. Light blue, light orange, and light green backgrounds identify the intervals 0–5, 12.5–17.5, and 25–30 seconds, respectively. Gaps marked by double slashes indicate omitted intervals; the displayed segments are not temporally contiguous. Case selection and annotated MAEs use the complete 30-second forecast. Colors, fixed-draw evaluation, and within-column amplitude scaling follow Figure 13.

does not imply the lowest MAE: Phase-Aligned Median Beat achieves an MAE of 0.810, compared with 1.161 for our model. Case selection follows the composite criterion, whereas the displayed MAE reports pointwise waveform error.

The failure cases exhibit mismatches in peak timing, amplitude, and the inter-beat waveform. In portions of the second failure case, our model generates more regular oscillatory structure than is supported by the observed signal. PatchTST and PSSM again exhibit amplitude attenuation, while the irregular fluctuations generated by the latent diffusion baselines do not consistently match the observed cardiac complexes. These examples distinguish generating structured waveforms from accurately predicting the future trajectory that actually occurs.

60s horizon. As shown in Figure 15, the favorable examples show that our model can preserve prominent cardiac complexes across the three displayed segments, although local timing and amplitude errors remain. In the first favorable example, PSSM achieves a full-horizon MAE of 1.545, lower than our model's 1.692, but substantially attenuates the observed cardiac complexes. This directly illustrates that lower pointwise error can coexist with weaker preservation of beat structure.

The failure cases show that a generated waveform can depart substantially from the observed future even when it retains oscillatory or beat-like structure. In the second failure case, our model achieves an MAE of 0.983, compared with 0.767 for PSSM; however, PSSM's prediction is nearly flat over substantial portions of the displayed intervals. This comparison illustrates two distinct limitations: attenuation of physiological variation and failure to predict its timing and waveform accurately.

Discussion. These examples illustrate distinct modeling behaviors, including amplitude attenuation, temporal misalignment, morphological mismatch, and preservation of beat-like structure despite higher pointwise error. They complement the quantitative results by showing why waveform



Figure 15: Qualitative examples and failure cases at the 60-second forecast horizon. Rows denote methods and columns denote matched cases. Light blue, light orange, and light green backgrounds identify the intervals 0–5, 27.5–32.5, and 55–60 seconds, respectively. Double slashes mark omitted intervals. Case selection and annotated MAEs use the complete 60-second forecast. Colors, fixed-draw evaluation, and within-column amplitude scaling follow Figure 13.

error, R-peak alignment, and physiological structure should be evaluated jointly. The examples are selected to visualize representative favorable and failure modes, while aggregate comparisons are based on the full test set.

A.19 Limitations and Implications

Limitations. This study uses Lead II ECG from a single emergency-care cohort. Generalization across institutions and to multilead settings remains to be established, while the small clinical weak-label subgroups limit the generalizability of subgroup findings. Future work will prioritize external validation, improved long-horizon dynamics modeling, and uncertainty evaluation with larger sampling budgets. ECG-WorldBench currently evaluates observation-conditioned waveform modeling and does not support counterfactual or action-conditioned simulation. Extending the benchmark with temporally aligned intervention information would enable comparison with intervention-aware models.

Implications. Our results show that pointwise error alone is insufficient for evaluating future ECG waveform modeling; temporal agreement, physiological structure, and semantic consistency provide complementary evidence. Ablation results further suggest that longer ECG histories and physiological supervision yield consistent, though generally modest, reductions in MAE and RMSE across all three horizons while also improving selected physiological and semantic measures. Model development should therefore balance waveform accuracy with physiological and representation-level consistency rather than optimizing a single objective. Grounded in continuous emergency ECG recordings, ECG-WorldBench supports the evaluation of physiological world models by examining whether they preserve patient-specific cardiac structure and capture temporal changes across future horizons.

A.20 AI USE STATEMENT

We employed large language models (LLMs) in a limited capacity, strictly for writing assistance tasks such as grammar correction, style refinement, and table formatting. The LLM was also used solely to generate code comments and explanatory notes in the supplementary implementation, and no core experimental code was produced by the LLM. All main model implementations, data processing pipelines, and training routines were entirely written and verified by the authors. All suggested edits were carefully reviewed and selectively incorporated by the authors. The scientific content, ideas, analyses, and conclusions presented in this paper are entirely our own. The authors take full responsibility for the paper's content, including any remaining errors or inaccuracies.